МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
"ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ"

Матеріали

III Всеукраїнської науково-практичної конференції

"ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ ТА ПРИКЛАДНА ЛІНГВІСТИКА"

ХАРКІВ 2014

УДК 004.9:81

Матеріали ІІІ Всеукраїнської науково-практичної конференції "Інтелектуальні системи та прикладна лінгвістика".

Харків, 17 квітня 2014 р. : матеріали конференції. – Харків: Національний технічний університет "Харківський політехнічний інститут", 2014. – 78 с.

В матеріалах розглядаються проблеми та перспективи розвитку інтелектуальних комп’ютерних систем та різних галузей прикладної лінгвістики, а саме корпусної лінгвістики, комп'ютерної лексикографії, машинного перекладу, лінгвістики Інтернету; питання використання інформаційних технологій в лінгвістиці, з метою дослідження та обробки мови.

ЗМІСТ

ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ ИДЕНТИФИКАЦИИ ЗНАНИЙ В ИНФОРМАЦИОННЫХ СИСТЕМАХ

Шаронова Н.В.
Национальный технический университет "Харьковский политехнический институт", г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60, e-mail: nvsharonova@mail.ru

На кафедре интеллектуальных компьютерных систем НТУ «ХПИ» проводятся исследования, целью которых является разработка методов и моделей для интеллектуальной обработки знаний, содержащихся в информационных системах широкого назначения, создание условий для внедрения, поддержания эффективного функционирования и развития самых разных информационных систем. При этом возникает вопрос: а что же такое знания?

В процессе работы над созданием современных компьютерных систем, решающих интеллектуальные задачи (в частности, понимание текстов на естественном языке), на первый план выдвигается проблема представления и извлечения знаний [1-4]. Основой функционирования таких систем является база знаний (БЗ), в которой хранятся сведения о заданной предметной области. Термин «знания» трактуется разными науками по-разному. Поскольку понятие «знания» достаточно сложно, то каждая наука, оперирующая этим понятием, вводит свои частичные его определения, и чаще всего эти определения носят антропоцентрический характер. Эти определения адресованы, прежде всего, человеку, а для построения формальных моделей они неприменимы. Существующее стандартизованное определение понятия «знания» выглядит следующим образом: «Знания – совокупность фактов, отношений, закономерностей и эвристических правил, отображающая уровень осведомленности о проблемах некоторой предметной области» [3].

Приведем еще одно полезное определение. *Знание – это результат адекватного отражения действительности человеком в виде представлений, понятий, теорий, суждений* [2]. Если говорить о знаниях, которые хранятся и обрабатываются компьютерными системами, то в этом случае определение понятия «знания» может быть сформулировано следующим образом: «*Знаниями принято называть хранимую (в ЭВМ) информацию, формализованную в соответствии с определенными структурными правилами, которую ЭВМ может использовать при решении проблем по таким алгоритмам как логические выводы*» [5].

Из различных определений понятия «знания», ориентированных на использование его при работе с компьютером, можно привести одно из самых удачных*: Знания – это информация, представленная в ЭВМ, которая имеет ряд особенностей: внутренняя интерпретируемость; структурированность; связность; семантическая метрика; активность.*

Если говорить о строго формализованном определении понятия «знания», то в работе [3] приводится следующее определение: «*Знание о факте – это отношение, выраженное некоторым высказыванием. Факт – это действительное состояние всех интересующих нас мест некоторого предметного пространства*». В этой же работе указывается, что знание не такое определенное понятие, как факт. Оно лишь ограничивает множество возможных состояний мест предметного пространства.

На уровне представления знаний в компьютерной системе отражены как отдельные элементы знаний, так и связи между ними. Уровень представления знаний отличается следующими особенностями: интерпретируемостью, наличием классифицирующих связей, наличием ситуативных отношений [3]. Кроме того, для уровня представления знаний характерны такие признаки, как наличие специальных процедур: обобщение, наполнение имеющихся в системе знаний и т.д. Рассмотрение эволюции проблемы машинного понимания в искусственном интеллекте обнаруживает, что именно разные типы знаний становились краеугольным камнем, методологическим фундаментом компьютерных понимающих систем нескольких поколений.

В недавно опубликованной монографии [1] проведен краткий анализ состояния проблемы в области моделирования идентификации знаний, проанализирована проблема формализации текстов, представленных на естественном языке. Рассмотрены методы интеллектуальной обработки информации: Data Mining и Text Mining. Показана необходимость дальнейшей разработки и исследования математического аппарата алгебры конечных предикатов, метода компараторной идентификации и возможности их применения в моделях представления знаний, системах искусственного интеллекта для формализации различных информационных процессов. Проанализировано и исследовано математическое и лингвистическое обеспечение автоматизированных информационных библиотечных систем. Приведен анализ систем обработки знаний с использованием онтологий.

Список литературы

1. Лингвотехнологии идентификации знаний в информационных системах : монография / О. В. Канищева, Н. В. Шаронова. – Saarbrücken, Deutschland : LAP LAMBERT Academic Publishing, 2013. – 173 с. – На рус. яз.

2. *Булкин В.И., Шаронова Н.В*. Математические модели знаний и их реализация с помощью алгебропредикатних структур: Монография, НТУ «ХПИ», МЭГИ, Харьков, Донецк, 2010, – 304 с.

3. *Бондаренко М. Ф.* Мозгоподобные структуры: Справочное пособие. / М. Ф. Бондаренко, Ю. П. Шабанов-Кушнаренко. Том первый. Под редакцией акад. НАН Украины И.В. Сергиенко. – К. : Наукова думка, 2011. – 460 с.

4. *Хайрова Н.Ф., Шаронова Н.В.* Лингвистические технологии экстракции и идентификации знаний // Тези доповідей Міжнародної науково-технічної конференції "Інтелектуальні технології лінгвістичного аналізу" (м. Київ, 22-23 жовтня 2013 р.). – К. : НАУ, 2013. – С. 7.

5. *Осуга С.* Обработка знаний / С. Осуга – М. : Мир, 1989. – 293 с.

ИССЛЕДОВАНИЕ СЕМАНТИКИ СЛОЖНОЙ ЯЗЫКОВОЙ СИСТЕМЫ КАК МЕЖДИСЦИПЛИНАРНОЙ ОБЛАСТИ СИСТЕМНО-КИБЕРНЕТИЧЕСКИХ ЗНАНИЙ

Хайрова Н. Ф.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: *nina_khajrova@yahoo.com*

В узком смысле прикладная лингвистика представляет собой ту часть языкознания, которая может быть применена при решении некоторой практической задачи. В частности, компьютерная лингвистика обхватывает область практической задачи автоматической обработки информации, представленной на естественном языке.

Основной задачей компьютерной лингвистики на сегодня остается проблема моделирования процесса понимания смысла текстов (перехода от текста к формализованному представлению его смысла) и проблема синтеза речи (перехода от формализованного представления смысла к текстам на естественном языке) [1]. Для решения данной задачи необходимо моделировать функции человеческого интеллекта, связанные с использованием естественного языка. К такого рода деятельности относятся: реферирование, перевод, экстракция и идентификация знаний, ответы на общие и специальные вопросы, перифраз и другие функции, связанные с пониманием текста или речи.

Для того чтобы формализовать, моделировать и автоматизировать такого рода деятельность необходимо формализовать естественный язык, рассмотрев его как цельную систему, включающую подсистемы, функции, структуры, формы и т.д.

Использование традиционных (статистических) подходов при формализации и решении задачи автоматической обработки семантики текстов на естественном языке в настоящее время становится малоэффективным. Практически единственным направлением исследований, прогрессивно решающим данную задачу, является использование при обработке языка моделей и методов теории интеллекта.

Системам, частично автоматизирующим одну из самых сложных функций человеческого интеллекта – понимание текста или речи, должны быть присущи следующие характеристики интеллектуальных систем: умение решать сложные плохо формализуемые задачи, способность к самообучению, развитые коммуникативные способности, адаптивность [2]. Разработка подобных систем возможна только при тесной взаимосвязи процедур таких областей системно-кибернетических знаний как искусственный интеллект и прикладная и компьютерная лингвистика. Направлениями взаимного влияния являются модели и технологии обработки информации, формальные языки и грамматики, модели и методы теории интеллекта, машинное обучение, лингвистические технологии и базы данных, лексикографические системы (рис. 1).

Рис. 1 – Определение области исследования семантики языковой системы как междисциплинарной области системно-кибернетических знаний.

Важнейшую роль в интеграции указанных научных направлений, по нашему мнению, должна сыграть концепция системного (комплексного) моделирования семантики сложной языковой системы, основанная на лингвистических технологиях, идеях и методах искусственного интеллекта.

Однако, для решения задачи формального представления семантического содержания лингвистических единиц необходимо разработать теоретическую базу, имеющую междисциплинарный характер и основанную на результатах, полученных в классической теории общего и прикладного языкознания, в компьютерной лингвистике, в ИИ, в теории систем и системном анализе [3].

Проблемы, связанные с анализом естественно-языковых объектов традиционно относят к области искусственного интеллекта. Фундаментальный вклад в развитие и становление методологической основы подходов и методов искусственного интеллекта, связанных с понимание и обработкой текстов на естественном языке, внесли выдающиеся отечественные и зарубежные ученые, развившие базовые элементы таких научных направлений, как теория автоматов, теория алгоритмов, математическая логика, теория программирования, алгебра конечных предикатов, компараторная идентификация [4, 5].

Все перечисленные направления, хотя и имеют глубокие проработки в своих исследованиях, не в состоянии обеспечить методологическим аппаратом процессы семантической разработки приложений лингвистического процессора, способные осуществлять обработку языка на уровне сравнимым с интеллектуальной деятельностью человека. Объективные препятствия, возникающие на пути анализа языковой системы, не позволяют

удовлетворительно решать проблему автоматизации его семантического анализа. Такое положение дел имеет место как из-за невозможности учета в настоящий момент всей специфической сложности системы семантики естественного языка, так и из-за отсутствия единой концептуальной основы построения информационной системы АОТЕЯ, принципиально учитывающей заданный уровень адекватности формализмов, моделей и методов явлениям и процессам языка.

С другой стороны, уже сформировались и активно развиваются качественно новые составляющие интеллектуальных информационных технологий, обещающие решить проблему включения в них формальных аппаратов традиционной математики (вычислительной алгебры, теории множеств, логико-алгебраические модели и др.) [6, 7], базирующиеся на аппарате знаний и связанных с ними моделях представления знаний.

Список литературы

1. Белоногов Г. Г. Компьютерная лингвистика и перспективные компьютерные технологии: моногр. / Г. Г. Белоногов, Ю. П. Калинин, А. А. Хорошилов. — М.: Рус. мир, 2004. — 248 с.

2. Кузнецова В. Л. Самоорганизация в технических системах: моногр. / В. Л. Кузнецова, М. А. Раков. — Киев: Наук. думка, 1987. — 200 с.

3. Згуровский М. З. Системный анализ: проблемы, методология, приложения / М. З. Згуровский, Н. Д. Панкратова — Киев: Наук. Думка, 2011. — 728 с.

4. Хомский Н. Три модели описания языка/ Н. Хомский // Кибернетический сборник: сб. переводов под ред. А. А. Ляпунова и О. Б. Лупанова. — Вып. 2. — М.: Иностр. лит., 1961. — С. 237—266.

5 Шабанов-Кушнаренко Ю. П. Компараторная идентификация лингвистических объектов: моногр. / Ю. П. Шабанов-Кушнаренко, Н. В. Шаронова — Киев: ИСДО, 1993. — 116 с.

6. Alejandro De Santos, Pedro G. Guillen, Eduardo Villa, Francisco Serradilla. Semantic Construction of Univocal Language; Information Theories and Applications, Vol.19, Number 3, 2012. — P. 211—215.

7. Dieter Jungnickel. Graph, Networks and Algorithms. Algorithms and Computation in mathematics. Volume5. – Springer Berlin Heidelberg New York, 2008. — 650 p.

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА

Канищева О. В.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: *olya-kanisheva@rambler.ru*

В настоящее время существует достаточно много определений понятия информационные технологии. В данной работе мы будем придерживаться определения приведенного профессором И. Бекманом [1].

Информационная технология включает в себя такие компоненты, как информатика, компьютерные технологии, Интернет и Всемирная паутина, Веб-разработки, управление данными, добыча и хранение данных, базы данных, информационная архитектура, информационная безопасность, криптография, системная интеграция, искусственный интеллект и др. Здесь выделяют такие направления, как: теоретическая информатика, кибернетика, программирование, искусственный интеллект, информационные системы, вычислительная техника, информатика в обществе, информатика в природе, информация в науке и технике.

Для дальнейшего понимания места компьютерной лингвистики среди других научных направлений нам понадобятся определения терминов информатика и кибернетика.

Термином *информатика* обозначают совокупность дисциплин, изучающих свойства информации, а также способы представления, накопления, обработки и передачи информации с помощью технических средств (рис. 1).

Рис. 1. Этапы информационного процесса

Кибернетика – наука об общих закономерностях процессов управления и передачи информации в различных системах, будь то машины, живые организмы или общество.

В 40-е годы наряду с идеей об универсальности схем управления в кибернетике развиваются и другие идеи: идея универсальной символики, идея логического исчисления, идея измерения информации через понятия вероятностной и статистической (термодинамической) теорий. В состав технической кибернетики входит теория автоматического управления, которая стала теоретическим фундаментом автоматики. Ведущее место в кибернетике занимает распознавание образов. Основная задача этой дисциплины – поиск решающих правил, с помощью которых можно было бы классифицировать многочисленные явления реальности, соотносить их с некоторыми эталонными классами.

Распознавание образов – это пограничная область между кибернетикой и искусственным интеллектом, ибо поиск решающих правил чаще всего осуществляется путём обучения, а обучение, конечно, интеллектуальная процедура.

Ещё одно научное направление связывает кибернетику с биологией. Аналогии между живыми и неживыми системами многие столетия волнуют учёных. Насколько принципы работы живых систем могут быть использованы в искусственных объектах? Ответ на этот вопрос ищет бионика – пограничная наука между кибернетикой и биологией. В свою очередь, нейрокибернетика пытается применить кибернетические модели в изучении структуры и действия нервных тканей. Недавно в кибернетике возникла – гомеостатика, изучающая равновесные (устойчивые) состояния сложных взаимодействующих систем различного типа. Это могут быть биологические системы, социальные системы, автоматические системы и др.

Математическая лингвистика занимается исследованием особенностей естественных языков, а также грамматик, позволяющих формализовать синтаксис и семантику таких языков. Это направление актуально в связи с развитием систем машинного перевода текстов с одних языков на другие.

Основным предметом математической лингвистики является разработка и изучение понятий, образующих основу формального аппарата для описания строения естественных языков. Возникновение математической лингвистики можно отнести приблизительно к 50-м гг. XX в., в связи с автоматизацией переработки языковой информации. Математическая лингвистика широко используются методы теории алгоритмов, теории автоматов и статистики [2]. Круг приложений математической лингвистики расширился – ее методы нашли применение в теории программирования.

Список литературы

1. Бекман И.Н. Информационные технологии и информатика [Электронный ресурс] : / И.Н. Бекман. – URL: http://profbeckman.narod.ru/InformLekc.files/Inf01.pdf

2. Математична лінгвістика [Текст] / В. В. Пасічник, Ю. М. Щербина, В. А. Висоцька, Т. В. Шестакевич. – Львів : «Новий світ – 2000», 2012. – 359 с.

ВИКОРИСТАННЯ МЕТОДІВ І МОДЕЛЕЙ ІНТЕЛЕКТУАЛЬНОЇ ОБРОБКИ НЕСТРУКТУРОВАНОЇ КРИМІНАЛІСТИЧНОЇ ІНФОРМАЦІЇ

Узлов Д.Ю.
Національний технічний університет
"Харківський політехнічний інститут",
м. Харків, вул. Пушкінська, 79/2, тел. 707–63–60,
e-mail: poputcik@mail.ru

Для того, щоб при роботі з неструктурованими або слабко структурованими масивами текстової інформації забезпечити працівника правоохоронних органів повною і релевантною інформацією, необхідно наблизити інформаційно-пошуковий запит до природної мови. Для цього в інформаційно-пошукових системах (ІПС), що працюють у різних предметних областях (наприклад, пошукові машини мережі Інтернет), використовуються інформаційно-пошукові мови (ІПМ) дескрипторного типу. Для використання мови дескрипторного типу при пошуку кримінально значимої інформації в системі необхідно створити криміналогічний тезаурус, який динамічно формується при зміні та коригуванні предметної області.

Усе вищезазначене обумовлює актуальність розвитку моделей та методів формалізації кримінально значимої інформації та застосування цих моделей для пошуку кримінально значимої інформації у неструктурованих або слабо структурованих текстових масивах, що власне і є напрямком досліджень автора.

У межах окресленої проблеми важливими є наукові завдання розробки моделей, методів, алгоритмів та програм, які здійснюють моделювання процесів інтелектуальної обробки інформаційних об’єктів з метою визначення їх основних характеристик для побудови інформаційного, математичного, лінгвістичного і програмного забезпечення ІПС.

Початковим етапом є відділення необхідної інформації за заздалегідь визначеними формальними ознаками та пошук інформації. Завданням інформаційного пошуку в рамках оперативно-розшукової діяльності є задоволення потреби в кримінально значимій інформації. Потреба правоохоронних органів, зокрема органів внутрішніх справ (ОВС), в інформації є досить різноманітною і визначається тактичною і стратегічною потребою розв'язуваної задачі. В основному, ці дані містяться в різних текстових масивах і не є чітко вираженою кримінальною інформацією. Так, наприклад, соціальні мережі, довідники, каталоги, форуми можуть містити дані про фігурантів кримінальної справи і при цьому можуть не мати кримінального забарвлення.

Таким чином, особливість вилучення інформації кримінальної значимості в основному визначається тим, що кримінальна значимість деякої множини даних буде визначатися динамічно тільки множиною метаданих. Множина метаданих формується в результаті опрацювання множини певних кримінальних даних, які містяться у відповідних інформаційно-криміналістичних базах знань.

Метою роботи є підвищення ефективності роботи сучасної інформаційної

криміналістичної системи за рахунок використання інтелектуальних методів і моделей обробки неструктурованої кримінально значимої інформації. Відповідно до зазначеної мети поставлено та вирішено такі задачі:

1) виконано аналіз методів і моделей ідентифікації і пошуку кримінально значимої інформації у неструктурованих текстових масивах при автоматизації інформаційних криміналістичних систем і сформульовано основні вимоги до розробки їхнього інформаційного, математичного та лінгвістичного забезпечення;

2) розроблено математичні та лінгвістичні засоби для розв'язання задач екстракції та обробки текстових масивів кримінально значимої інформації на основі моделювання лінгвістичної діяльності людини і інтелектуального аналізу даних методом компараторної ідентифікації;

3) розроблено засоби моделювання процедур екстракції та ідентифікації кримінально значимих фактів із текстів для застосування у задачах оперативно-розшукової діяльності;

4) удосконалено моделі процесів обробки кримінально значимої інформації за рахунок динамічного наповнювання об’єктно-орієнтованого тезауруса оперативно-розшукової діяльності у криміналістичних системах;

5) розроблено модель бази знань інформаційної криміналістичної системи на основі побудови лінгвістичного процесору, об’єктно-орієнтованого тезауруса та словника колокацій на базі семантичної мережі понять зі сталими зв’язками між ними та антиципаційного алгоритму;

6) виконано практичну реалізацію запропонованих методів і математичних моделей, впроваджено результати дисертаційної роботи у практику створення реальних інформаційних криміналістичних систем.

Об’єктом дослідження є кримінально значима інформація в автоматизованих інформаційних системах. Предметом дослідження є методи і моделі інтелектуальної обробки слабо структурованої кримінально значимої інформації.

Висновок. Комплекс розв’язаних задач сприяє розвитку моделей та методів формалізації кримінально значимої інформації та застосуванню цих моделей для пошуку кримінально значимої інформації у неструктурованих або слабо структурованих текстових масивах реальних ІПС.

Список літератури

1. *Бандурка О.М*. Особливості виділення кримінально значимої інформації в текстових масивах / О.М. Бандурка, М.М. Зацеркляний, Д.В. Лазарєв, Д.Ю. Узлов // Наше право, – ХНУВС, Кримінологічна асоціація України, МАУП та ін. – Харків, Спеціалізоване видавництво ЮНЕСКО, № 2 ч.1, 2011, с. 79-84.

2. *Узлов Д.Ю.* Модель извлечения криминально значимых данных из массивов неструктурированной информации / Д. Ю. Узлов, Н. Ф. Хайрова // Військова освіта і наука: сьогодення та майбутнє: тези доп. VII Міжнар. наук.-практ. конф., Київ. 24-25 лист. 2011р. — С.81—82.

3. *Бондаренко М. Ф.* Мозгоподобные структуры: Справочное пособие. / М. Ф. Бондаренко, Ю. П. Шабанов-Кушнаренко. Том первый. Под редакцией акад. НАН Украины И.В. Сергиенко. – К.: Наукова думка, 2011. – 460 с.

ПРИМЕНЕНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ ТЕХНОЛОГИЙ ЭКСТРАКЦИИ И ИДЕНТИФИКАЦИИ ЗНАНИЙ В КОРПОРАТИВНЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

Шкапо С.В., Аджит Пратап Сингх Гаутап
Национальный технический университет
"Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: nvsharonova@mail.ru

Целью проводимых на кафедре интеллектуальных компьютерных систем НТУ «ХПИ» исследований является разработка методов и моделей для интеллектуальной обработки знаний, содержащихся в информационных системах широкого назначения, создание условий для внедрения, поддержания эффективного функционирования и развития корпоративных информационных систем.

Под корпоративной средой знаний в современных информационных системах понимается комплекс методического, организационного, программного, информационного и технического видов обеспечения, нацеленных на достижение и поддержание в компании заданного уровня компетенции в избранной области. В соответствии с принятым в международных стандартах подходом, компетенция – это совокупность знаний, навыков и личностных характеристик сотрудников, которые: связаны с исполнением работы и оказывают на него существенное влияние; могут быть структурированы и измерены в соответствии с признанными стандартами; могут быть усовершенствованы посредством обучения и развития. Заданный уровень компетенции сотрудников компании является определяющим фактором при формировании и внедрении корпоративной среды.

В состав корпоративной среды знаний входят элементы, которые планируется рассмотреть с точки зрения математического моделирования основных процессов, происходящих в них: это корпоративная система диагностики знаний; электронная библиотека; корпоративный портал знаний; корпоративная система дистанционного и тренингового обучения; корпоративная справочная служба; корпоративная система наставничества; а также корпоративная система карьерного развития.

Знания, представляющие основу всех перечисленных элементов корпоративной среды, представлены преимущественно в текстовом виде, обработка их представляет сложную задачу обработки слабо формализованной и слабоструктурированной информации. Предполагается использование метода компараторной идентификации и средств алгебры логики для построения эффективных моделей экстракции и идентификации знаний в корпоративных системах.

Алгебра предикатов дает возможность описывать функции интеллекта в виде предикатных уравнений. В работах [1-4] показано, что, используя математический аппарат алгебры предикатов, можно описывать в виде уравнений ал-

гебры предикатов алфавитные операторы, первичные понятия конечной математики, основные понятия алгебры множеств и теории отношений. Средствами алгебры предикатов можно также моделировать булевы функции, булевы отношения, переключательные функции [3]. Уравнения алгебры предикатов позволяют описывать конечные математические структуры, вполне конечные автоматы [2]. С помощью алгебры предикатов можно моделировать слова и простейшие процессы обработки слов. На языке алгебры предикатов можно осуществить математическое описание алгоритмической деятельности человека, а также выражений и действий над ними. По формулам алгебры предикатов можно строить переключательные цепи, которые называют АП структурами.

Алгебра предикатов позволяет моделировать конечные математические структуры [2], которые имеют большое значение для теории интеллекта. Сущность моделирования интеллектуальных процессов заключается в том, что для каждого из этих процессов подыскивается соответствующая математическая структура, которая описывается в виде уравнений алгебры предикатов. Эти уравнения при необходимости реализуются аппаратно в виде устройств, которые воспроизводят моделируемые интеллектуальные процессы. Человек использует различные приемы для формирования математических структур. К ним относятся: формирование множеств из наличных элементов, формирование декартовых произведений множеств, образование подмножеств уже имеющихся множеств. Перечисленные приемы могут использоваться в различных комбинациях и многократно. Полученные структуры могут быть использованы для построения других структур.

Алгебра предикатов позволяет описывать отношения, зафиксированные в текстах естественного языка. В последних работах, развивающих теорию интеллекта [2, 3] вводятся понятия линейного логического оператора как инструмента решения агебропредикатных уравнений и бинарных логических сетей, которые представляют собой графическое представление результата бинарной декомпозиции многоместного предиката.

Список литературы

1. *Булкин В.И., Шаронова Н.В*. Математические модели знаний и их реализация с помощью алгебропредикатних структур: Монография, НТУ «ХПИ», МЭГИ, Харьков, Донецк, 2010, – 304 с.

2. Лингвотехнологии идентификации знаний в информационных системах : монография / О. В. Канищева, Н. В. Шаронова. – Saarbrücken, Deutschland : LAP LAMBERT Academic Publishing, 2013. – 173 с. – На рус. яз.

3. *Бондаренко М. Ф.* Мозгоподобные структуры: Справочное пособие. / М. Ф. Бондаренко, Ю. П. Шабанов-Кушнаренко. Том первый. Под редакцией акад. НАН Украины И.В. Сергиенко. – К. : Наукова думка, 2011. – 460 с.

4. *Хайрова Н.Ф., Шаронова Н.В.* Лингвистические технологии экстракции и идентификации знаний // Тези доповідей Міжнародної науково-технічної конференції "Інтелектуальні технології лінгвістичного аналізу" (м. Київ, 22-23 жовтня 2013 р.). – К. : НАУ, 2013. – С. 7.

ИСПОЛЬЗОВАНИЕ ФОНОСЕМАНТИЧЕСКОЙ ОЦЕНКИ СЛОВ-МОДИФИКАТОРОВ В СЛОВАРЯХ ОЦЕНОЧНОЙ ЛЕКСИКИ

Игнатьев А.М.
Национальный университет гражданской защиты Украины,
г. Харьков, ул. Чернышевская, 94, тел. 050-733-78-33,
e-mail: Ignatiew@yandex.ru

За последние десять лет интерес к области анализа эмоциональной тональности текстов сильно возрос. Стоит отметить, что на текущем этапе развития в данной области существует много нерешенных проблем. Анализ эмоциональной окраски текста затруднителен не только в связи с проблемой выделения единиц оценки тональности, но и ввиду неоднозначности эмоциональной составляющей лексических компонент. Например, в рамках одной и той же предметной области "высокая стоимость" – отрицательный аспект товара, в то время как "высокое качество" – положительный. Таким образом, используемые методы тонального анализа предметно зависимы, т.е. для различных предметных областей необходимо составлять различные словари.

Основные подходы к определению тональности можно разделить на следующие категории [1]:

1. Подход, основанный на правилах (rule-based approach), заключается в применении набора правил, выявленного экспертами на основе анализа предметной области.

2. Подход, основанный на использовании словарей оценочной лексики (affective lexicons). Для каждого слова, встречаемого в документе, из словаря получают значение тональности. Чтобы получить итоговую тональность необходимо взять среднее арифметическое или вычислить сумму значений тональности всех слов из документа.

3. Подходы, основанные на обучении с учителем (supervised learning). Алгоритм классификации тренируется на основе обучающей выборки (корпуса), состоящей из документов, классы которых заранее известны.

4. Подходы, основанные на обучении без учителя (unsupervised learning). Отличие состоит в том, что в этом случае для тренировки алгоритма используется обучающая выборка, состоящая из документов, классы которых заранее неизвестны (или известны, но эта информация не используется алгоритмом).

В ходе экспериментов методы, основанные на словаре эмоциональной лексики, при решении задачи автоматической классификации текстов по тональности показали результаты, несколько превосходящие результаты метода опорных векторов (Support Vector Machine, SVM) и простейшего способа классификации (baseline) [2]. Исследования показали, что методы на основе словаря показывают достаточно неплохие результаты при классификации текстов.

Однако, кроме оценочных слов для выбранной предметной области, в текстах встречается множество слов-модификаторов, в зависимости от которых можно увеличивать или уменьшать вес следующего за ним оценочного слова. Все слова-модификаторы можно разделить на две группы в зависимости от их

направленности. К первой группе относятся слова-модификаторы, которые увеличивают эмоциональный вес соседнего слова (например, «особенно»), ко второй – те, которые уменьшают ее (например, «незначительно»). Для изменения веса следующего за модификатором слова можно использовать метод простого сложения и вычитания. Если модификатор увеличивает эмоциональный вес слова, то к его оценке можно добавлять фиксированное число, иначе – вычитать это же число.

Однако, недостатком данного подхода является то, что он не учитывает широкий диапазон модификаторов в пределах группы. Например, модификатор «абсолютно» очевидно сильнее изменяет эмоциональный вес слова, чем модификатор «значительный». Также при усилении слова с уже большим весом увеличение его эмоционального веса должно быть больше по сравнению со словом, обладающим меньшим весом. Например, «действительно восхитительный» и «действительно хороший». Возможно использовать подход, который в зависимости от слова-модификатора изменяет вес соседнего слова на некоторый процент. Например, если слово «хорошо» имеет вес 5, а модификатор «действительно» имеет относительную оценку 20%, то «действительно хорошо» будет иметь вес 5*(100% + 20%) = 5*1,2 = 6. В качестве модификаторов используются наречия и прилагательные. Такой подход был рассмотрен в работе [3], в которой процентные значения для слов-модификаторов фиксировались на основе экспертных оценок.

Нами предлагается поход, в котором процентные значения для слов-модификаторов будут вычисляться на основе их фоносемантических оценок по различным шкалам. Таких шкал может быть несколько («хороше-плохое», «быстрое-медленное» «сильное-слабое» и т.д.). Оценка каждого слова-модификатора вычисляется как среднее арифметическое всех его фоносемантических оценок по всем предложенным шкалам. Такой поход позволит учесть эмоциональную окраску самих слов-модификаторов и, как следствие, получить более точные веса анализируемых слов.

Предлагаемый способ оценки весов слов позволит избежать коллизий, возникающих при применении слов, выражающих отрицание, к словам-модификаторам. Простое инвертирование эмоционального веса оценочного слова хорошо работает лишь в некоторых случаях, но часто может привести к нежелательному результату. Например, «не очень хорошо» может оказаться более отрицательно, чем «плохо». В работе [3] вместо смены знака, значение эмоционального веса сдвигается к противоположной полярности на фиксированную величину. Нами предлагается учитывать отрицание изменением веса соседнего слова на процент его модификатора, взятого со знаком «минус».

Список литературы

1. *Pang B., Lee L.* Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval, v.2 n.1-2, January, 2008. - pp. 1-135.

2. Российский семинар по оценке методов информационного поиска (РОМИП). URL: http://romip.ru.

3. *Taboada M., Brooke J., Tofiloski M., Voll K., Stede M.* Lexicon-based methods for sentiment analysis // Computational Linguistics, 37(2): 2011. - pp. 267–307.

ВИКОРИСТАННЯ МЕТОДУ АВТОМАТИЧНОЇ ЕКСТРАКЦІЇ ВІДНОШЕНЬ СЕМАНТИЧНОЇ БЛИЗЬКОСТІ ДЛЯ РОЗРОБКИ БАЗ ЗНАНЬ

Петрасова С. В.
Національний технічний університет
«Харківський політехнічний інститут»,
м. Харків, вул. Пушкінська 79/2, тел. 707-63-60,
e-mail: svetapetrasova@gmail.com

Сучасний стан розробок баз знань та їх систематизації потребує впровадження універсальних інтелектуальних систем, важливим завданням яких є екстракція інформації з текстів та представлення її у вигляді формальної системи знань.

Найбільш перспективним способом формального вираження знань сьогодні є семантична мережа. Це обумовлено, передусім, наочністю представлення знань та можливістю в явному вигляді виражати семантичні відношення між поняттями [1].

Але розповсюдження даного способу представлення знань стримується як неоднозначністю вираження знань на природній мові, так і трудомісткістю та складністю розробки семантичної мережі.

В такому разі одним з найбільш повних джерел знань для автоматичної побудови бази знань можуть слугувати такі універсальні засоби представлення, накопичення та передачі інформації, як тексти. Серед усіх текстових джерел саме глосарії представляють тексти природної мови з найбільш концентрованим смисловим навантаженням.

В даному дослідженні пропонується використання методу автоматичної екстракції відношень семантично близьких понять для розробки семантичних мереж, який ґрунтується на знаннях глосарія, виражених дефініціями термінів даних об’єктів [2].

Для побудови логічної схеми виявлення семантично близьких термінів вводиться метричний простір лінгвістичних смислових одиниць Θ, який визначається як множина лінгвістичних одиниць лексикону T, на якому граматичні правила задають відношення між одиницями, що виступають обмеженнями для коректних синтаксичних структур [3].

Міру семантичної близькості f формально визначимо співвідношенням через відповідні дефініції глосаріїв d_1 та d_2 як потужності множин, утворених теоретико-множинним перетином та об’єднанням множин термінів дефініцій:

$$f(t',t'') = \frac{2\,|\,d_1 \bigcap d_2\,|}{|\,d_1\,| + |\,d_2\,|},$$

де $d_1 \bigcap d_2$ – спільні терміни дефініцій, а $|d_1|+|d_2|$ – всі терміни дефініцій d_1 и d_2.

В створеному просторі концептів з одним і тим самим сигніфікативним смислом можна виявити такі категорії семантичних відношень, як приналежність до класу, гіперонімія, гіпонімія та меронімія. Для формалізації описаних типів відношень застосовуються шаблони лексичних послідовностей:

$$NN_1 \rightarrow \mathrm{Re}\, l_z \rightarrow NN_2,$$

де NN_1 и NN_2 – зв'язані концепти, представленні ключовими словами та словосполученнями глосаріїв, $\mathrm{Re}\, l$ – лексичні ланцюжки, які виражають відношення z (табл. 1).

Таблиця 1 – Приклади шаблонів лексичних послідовностей

Семантичне відношення, z	**Лексичні ланцюжки,** $\mathrm{Re}\, l$
Приналежність до класу	"є", "вважається"
Гіперонімія	"сукупність", "комплекс", "набір", "сімейство"
Гіпонімія	"наприклад", "тип", "(різно)вид", "екземпляр"
Меронімія	"частина", "елемент"

Таким чином, неоднозначність тлумачення та представлення природної мови є характерною особливістю текстових ресурсів, що не дозволяє однозначно формалізувати виявлення семантичних відношень з текстів. Для вирішення даної проблеми розглянуто метод автоматичної екстракції відношень семантичної близькості, який основується на використанні глосарія як природномовного тексту, що найбільш повно концентрує знання.

Список літератури

1. *Manning C. D.* Introduction to Information Retrieval / Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. – Cambridge University Press, England, 2009. – 544 p.

2. *Кобозева И. М.* Лингвистическая семантика: Учебное пособие. / И. М. Кобозева. – М. : Эдиториал УРСС, 2000. – 352 с.

3. *Хайрова Н. Ф.* Определение семантической близости на основе когнитивного подхода. / Н. Ф. Хайрова, Н. В. Шаронова, Н. В. Борисова // Бионика интеллекта: науч.-техн. журнал, 2013.

ЕЛЕМЕНТИ ДИСТАНЦІЙНОЇ ОСВІТИ В НАВЧАННІ ІНОЗЕМНИМ МОВАМ
(НА ПРИКЛАДІ ВИКЛАДАННЯ АНГЛОМОВНОГО ЛЕКСИЧНОГО МАТЕРІАЛУ)

Рижкова В.В.
Національний аерокосмічний університет
ім. М.Є. Жуковського «Харківський авіаційний інститут»
м. Харків, вул.. Чкалова, 17, тел. 788-47-06,
e-mail: foxenfoxen@mail.ru

Придбання нових знань та навичок, практично корисних у роботі в епоху інформаційного суспільства, значно розширює можливості самореалізації й сприяє кар'єрному росту. В цьому аспекті дана робота видається *актуальною*, оскільки висвітлює перспективи та нові можливості дистанційного навчання для студентів-філологів.

Дослідження проводиться в площині залучення комп'ютерних технологій до уведення граматичного, лексичного, текстового, відео та аудіо англомовного матеріалу студентам, які вивчають англійську мову.

Практична значимість дослідження полягає в тому, що створені масиви тестів та мультимедійного супроводу можуть застосовуватися у роботі зі студентами дистанційної та денної форм навчання у рамках проведення занять, для перевірки загальної мовної компетенції студентів, для забезпечення завдань на самостійне опрацювання студентами.

Відсутність спільного підходу до розробки методики практичних курсів дистанційного навчання іноземним мовам можна розглядати як основну методичну проблему у цій галузі. Невирішеність цієї проблеми, яка помітно гальмує процес упровадження дистанційного навчання іноземним мовам до практики освіти, можна пояснити як її відносною новизною, так і її складністю. Це зумовлено тим, що суть проблеми знаходиться у точці перетину двох предметних галузей. Перша — новітні інформаційні технології, друга — власне методика навчання іноземним мовам.

Кафедра прикладної лінгвістики Національного аерокосмічного університету ім. М.Є.Жуковського «ХАІ» розробила та, із застосуванням системи Moodle за допомогою лабораторії дистанційної освіти ХАІ, розмістила електронні підручники з дисциплін навчального плану, що пропонуються студентам спеціальності «Прикладна лінгвістика» заочної форми навчання. Ми наводимо приклад частини електронного підручника (навчання лексиці за певною тематикою) з англійської мови (1 курс 1 семестр), адже успішне оволодіння іншомовним лексичним матеріалом – одна з найважливіших умов засвоєння мови [1, с. 109]. Оскільки для вивчення іноземної мови необхідні знання з граматики та лексики, а також навички в аудіюванні, читанні та письмі, розроблений електронний підручник з англійської мови має блоки: **Grammar, Speech, Listening, Reading, Writing** із наданням теорії, тестами та завданнями.

Інформаційна насиченість сучасного світу вимагає спеціальної підготовки навчального матеріалу до його пред'явлення, щоб у візуально доступному для огляду вигляді надати студентам основні або необхідні відомості [2, с. 135]. Блок «Speech» із електронного підручника з англійської мови за темою *My Working Day* в системі Moodle дистанційної лабораторії «ХАІ» має наступний вигляд.

ТЕКСТ 3: **My Working Day** (навчальний текст та звуковий супровід**)**
My Working Day – навчальна презентація в Power Point
My Working Day – TEST-презентація в Power Point (30 завдань)
ПЕРЕВІР СЕБЕ:
ТЕСТОВІ ЗАВДАННЯ ЗА ТЕМОЮ **My Working Day**
ТЕСТ 1.1. Read the definition and decide whether the statement is True or False (50 шт.)
ТЕСТ 1.2. Complete the sentences (50 шт.)
ТЕСТ 1.3. Translate into Russian (50 шт.)
ТЕСТ 1.4. Translate into English (50 шт.)
ТЕСТ 1.5. Find pairs of synonyms or antonyms (100 шт: 5 завдань по 20 слів)

Навчальна тема подається у вигляді електронного тексту, що має звуковий супровід, з переліком слів після нього. Для зручності у створенні навчальної презентації, що покликана полегшити процес засвоєння студентами нового лексичного матеріалу, було установлено асоціативні зв'язки форми та зображення. Для розроблення навчальних мультимедіа презентацій за розмовними темами користувалися найпоширенішою в Україні російськомовною версією PowerPoint, яка входить до складу інтегрованого пакета MSOffice. Для запису звукового супроводу було застосовано програму Adobe Audition 1.5. Диктора англійських слів (до усіх слів теми, що вивчається, дібрано картинку – зоровий образ лексичної одиниці – та звуковий супровід) було обрано з Google Translator. Він найбільш точно і правильно вимовляє слова, що дуже важливо для студентів. Таким чином, у студента є можливість побачити графічне оформлення лексичної одиниці та словосполучення, зоровий образ та почути їх.

Далі з метою зацікавлення студентів було створено PowerPoint TEST- презентацію на 30 завдань для перевірки вивченого студентами матеріалу. Після знайомства з темою та перевірки вивченого у форматі PowerPoint TEST- презентацій починається дуже детальне опрацювання теми із завданнями різного характеру, як наприклад: дати визначення поняттю, що зазначається у тексті; вставити пропущене слово, яке підходить за змістом, у речення; знайти синонім/антонім до слова; перекласти речення з англійської на українську\російську мову та навпаки.

Варто зазначити, що для кращого засвоєння матеріалу та перевірки здобутих знань доцільним є виконання тестів відкритого типу, оскільки саме такі тести виключають момент вгадування та дають можливість оцінити об'єктивно знання кожного студента.

Кожний тест був ретельно опрацьований щодо наповнення лексичним та граматичним матеріалом та пройшов попередню апробацію в групах студентів денної форми навчання. Перевірка виконаних завдань (окрім перекладу) прово-

диться комп’ютером. Програма дає можливість повторного проходження тесту, якщо студент не набрав необхідну кількість балів (мінімум 65 % від усього тесту), кожне завдання оцінюється у 2 бали.

Впровадження мультимедійних технологій у навчальний процес ВНЗ дозволяє підвищити якість знань, посилити мотиваційний аспект, а на цій основі – пізнавальний інтерес у студентів до підвищення рівня фахової підготовки.

Список літератури

1. Артемчик Г. Про сучасні підходи до вивчення і викладання іноземних мов / Г. Артемчик. — К. : Рідна школа, 2003. – 159 с.

2. Клокар Н. Методологічні основи запровадження дистанційного навчання в системі підвищення кваліфікації / Н. Клокар. — М. : Шлях освіти, 2007. — 167 с.

АВТОМАТИЗАЦІЯ ПОБУДОВИ ПРЕДМЕТНОЇ ГАЛУЗІ КУРСІВ У РАМКАХ СИСТЕМИ ДИСТАНЦІЙНОГО НАВЧАННЯ

Кузіков Б.О.
Сумський державний університет,
м.Суми, вул. Римського-Корсакова, 2, тел. (0542) 77–08–27,
e-mail:b.kuzikov@dl.sumdu.edu.ua

Одним з ключових компонентів моделі адаптивної системи дистанційного навчання (аСДН) є модель предметної галузі. Структура моделі та якість наповнення цього компоненту є важливими при побудові аСДН на основі СДН, що не має таких властивостей, бо обумовлюють перелік підходів до адаптації, які можуть бути застосовані. Тому в рамках вирішення задачі побудови аСДН на основі СДН СумДУ була розроблена модель предметної галузі навчальних курсів та реалізовано сервіс імпорту нових та адаптації існуючих матеріалів.

В рамках дослідження модель предметної області представлена множиною типізованих понять. Поняття курсу поділяються на загальнонаукові та предметно-орієнтовані. Між поняттями можуть встановлюватись відношення синонімії. Інші типи зв'язків не передбачені.

Контент у СДН СумДУ представлено набором пов'язаних гіпертекстових об'єктів. Для зв'язування понять та окремих навчальних об'єктів запроваджено спеціалізований сервіс. Вважається, що один об'єкт може бути проіндексований кількома поняттями. Виходячи з гіпотези про логічність впорядкованості навчальних матеріалів курсу сервіс дозволяє спрогнозувати роль поняття в об'єкті: об'єкт потребує попереднього ознайомлення з поняттям чи поняття вводиться у ньому вперше. При підготовці до використання в аСДН під час завантаження навчальних об'єктів проводиться їх аналіз. Алгоритм роботи сервісу «Аналіз документів» представлений на рис. 1. Для аналізу використовуються документи у форматі html. Якщо документ представлений в інших форматах (odt, doc), попередньо застосовується сервіс перетворення документів.

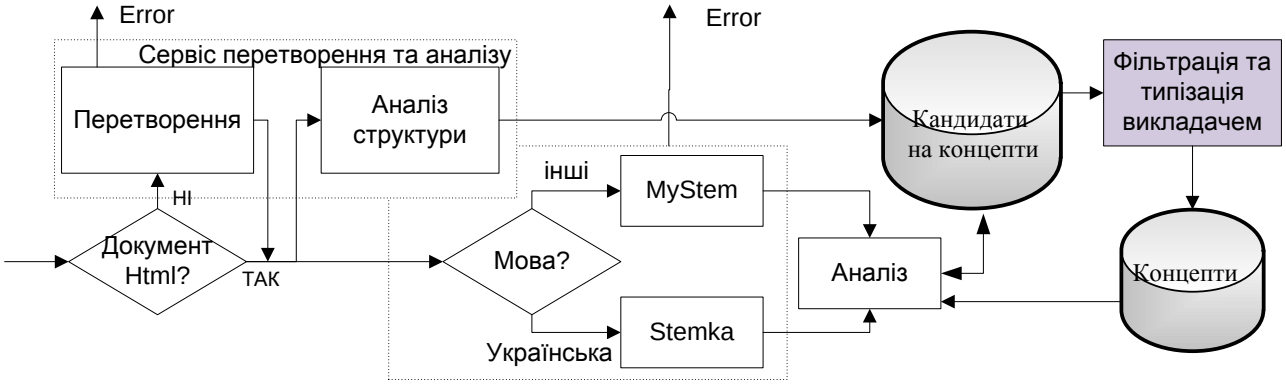


Рис. 1.– Структурна схема сервісу аналізу документів

Аналіз документа проводиться у два етапи. По-перше, на основі структурних особливостей документа сервіс аналізу структури намагається виділити в документі визначення або ключові слова [1]. Оптимальним у цьому випадку є варіант, коли документ доповнено метаінформацією щодо ключових слів. Такі

ключові слова позначаються як вихідні поняття. На другому етапі проводиться аналіз документа на основі морфологічного розбору та виділення кандидатів на ключові слова, використовуючи базу відомих понять. При цьому використовуються модулі сторонніх розробників MyStem [2] та Stemka [3]. Поняття, виділені через перелік ключових слів та на основі аналізу документа, заносяться в таблицю кандидатів. Аналізуючи таблицю кандидатів, автор навчального об'єкта може уточнити тип поняття (предметно-орієнтоване чи загальнонаукове) та його роль (поняття є базовим чи результуючим для тексту, що аналізується).

Поняття, його тип та зв'язки із документом зберігаються у глобальному сховищі. Схема частини бази даних, що відповідає за збереження понять, представлена на рис. 2. Урахування синонімії понять реалізовано через зв'язок parent_id→term_id у таблиці Terms.Term. Сховище кандидатів реалізовано виділенням частки таблиці (partition) Terms.TermUsage за критерієм (is_visible = false).

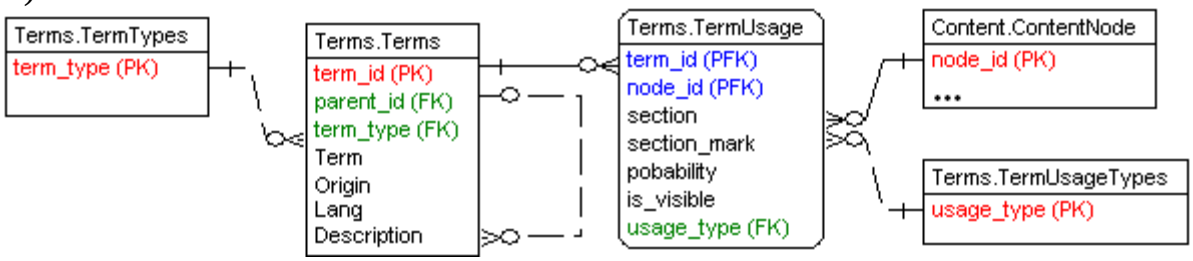


Рис. 2.– Схема таблиць БД, пов'язаних із збереженням понять

Первинне наповнення глобального сховища понять було виконано за допомогою розбору глосаріїв курсів. Це гарантувало, що терміни, які були заздалегідь виділені автором, обов'язково потраплять до онтології понять дисципліни, що обробляється. Після розбору окремих документів проводиться формальна перевірка повноти курсу із застосуванням авторського підходу [4].

Запроваджений сервіс використовується при розробці навчальних курсів у СДН СумДУ та є органічним доповненням нового модуля розробки навчальних курсів, у якому провідну роль відведено самим авторам курсів. Подальшими напрямками роботи є підвищення точності виділення понять та розширення переліку форматів, що підтримуються.

Список літератури

1. Кузиков Б. О. Использование Libre Office в дистанционном обучении [Текст] / Б. О. Кузиков // Міжнародна науково-методична конференція «Якість вищої освіти : методологічні та методичні підходи щодо впровадження дистанційних технологій навчання», 23-24 січ. 2013 р., м. Полтава. – Полтава, 2013. – Ч. 2. – С. 112-114.

2. Segalovich Ilya. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine [Електронний ресурс] / Segalovich Ilya. - Режим доступу : http://download.yandex.ru/company/iseg-las-vegas.pdf

3. Коваленко А. Вероятностный морфологический анализатор русского и украинского языков [Електронний ресурс] / А. Коваленко // Системный администратор. - 2002. – № 1. - Режим доступу : http://samag.ru/archive/article/47

4. Kuzikov B. Using semantic web and covering context by test for course formal testing [Text] / V. Lubchack, B. Kuzikov, K. Kirichenko // 8th Int. Conference on Emerging eLearning Technologies and Applications, High Tatras, Slovakia. – 2010. – C. 135-140.

НЕКОТОРЫЕ АСПЕКТЫ ОСУЩЕСТВЛЕНИЯ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕКСТОВ НА ОСНОВЕ РАСПОЗНАВАНИЯ ИХ АДРЕСАТОВ

Глазкова А.В.
Тюменский государственный университет, г. Тюмень, ул. Семакова, д. 10, +79091826371, anya_kr@aol.com

1. Введение

В связи с активным развитием технологий разработки интеллектуальных систем в настоящее время увеличивается потребность в исследованиях, направленных на улучшение механизмов информационного поиска. Использование поисковых систем, электронных библиотек, спам-фильтров немыслимо без применения инструментов обработки текстовой информации.

Данная работа направлена на рассмотрение вопроса автоматического определения адресата текста – в частности, возможности классификации текстов в зависимости от того, какой возрастной аудитории они адресованы. Мы хотим обрисовать определенный минимум, необходимый для создания системы, реализующей заявленную функцию. В дальнейшем планируется определить состав набора признаков для классификации, сделать предложения по созданию базы знаний и обучающего корпуса, методам классификации и лингвистического анализа.

2. Задача классификации и ее формальная постановка

Задача классификации текстов заключается в определении принадлежности текста одному или нескольким классам. Для каждого документа-объекта при этом выделяются наборы признаков – слов и их взаимозависимых наборов. Для формирования наборов этих признаков для каждого документа используются лингвистические и статистические методы [1]. Численные значения, принимаемые объектами того или иного класса, вычисляются в процессе обучения классификатора. По завершении обучения принадлежность текста к классу определяется при помощи проведения анализа признаков текста с учетом полученных весовых значений.

Автоматическая классификация может применяться в таких областях информационного поиска:

- поиск в электронных библиотеках и сети Интернет;
- фильтрация почтового спама;
- составление интернет-каталогов;
- подбор контекстной рекламы;
- снятие неоднозначности при автоматическом переводе текстов и др.

Приведем формальную постановку задачи классификации. Пусть дано множество категорий и множество документов . Целевая функция , которая для каждой пары <документ, категория> определяет, соответствуют ли они друг другу, неизвестна. Задача состоит в построении классификатора , максимального близкого к функции [2]. Основными подходами к решению данной

задачи являются наивный байесовский подход, метод k ближайших соседей, построение деревьев решений, использование метода опорных векторов и создание нейронных сетей [3].

3. Выявление характеристик адресата текста

В рамках разработки методов и алгоритмов автоматической классификации текстов рассматриваются вопросы распределения текстов по жанрам, времени написания, автоматического распознавания автора и языка. Любой текст, как известно, явно или не явно предназначается конкретному читателю как в широком смысле – например, группе людей, говорящих на определенном языке, так и в узком – например, представителям одной возрастной категории [4]. При этом текст как бы включает в себя образ «своей» идеальной аудитории, аудитория – «своего» текста [5].

В рамках своей коммуникативной деятельности автор составляет текст, имея установку на максимально полное доведение до адресата своего замысла для того, чтобы адресат его (автора) понял. Речь должна быть ориентирована на слушателя, и естественным следствием такой установки является намерение автора использовать такие содержание и структуру прогнозируемого текста, а также такие средства языка для их выражения, которые в своей совокупности были бы доступны пониманию реципиента, которому адресован текст. В работе Каменской О.Л. [6] рассматривается понятие коммуникативного портрета адресата текста и в связи с этим выделяются основные «слагаемые» личности реципиента, необходимые для понимания адресованного ему текста. К ним относятся:

- индивидуальное знание адресата в той области, в рамках которой будет протекать коммуникативный акт (то есть непрерывно кон-струируемая и модифицируемая динамическая система данных, которыми располагает индивид);
- наличие специальных знаний в области, которой посвящен текст;
- объем активного тезауруса личности в данной области знаний (под этим термином понимается организованное знание, которым обладает субъект о словах и других вербальных символах).

Таким образом, к составу набора признаков для автоматического распознавания адресата текста можно отнести данные, полученные на основе словарного состава документа – подобной характеристикой может быть, например, отношение количества терминов к общему числу слов.

С рассматриваемой задачей тесно связаны исследования удобочитаемости или читабельности документов (Readability), опирающиеся на анализ синтаксиса и словаря текста. Основными критериями, оказывающими воздействие на значение показателя удобочитаемости, считаются количество слов в предложении, количество терминов в тексте, число символов в слове [7]. Число и состав критериев может меняться в зависимости от жанра, коммуникативной задачи и языка текста [8]. В качестве иных особенностей, влияющих на классификацию, можно указать количество сложносочиненных и сложноподчиненных предложений, количество обособлений, причастных и деепричастных оборотов. Одной из наиболее часто применяемых мер определения сложности восприятия текста читателем, адаптированных для русского языка, является индекс Флэша.

Он вычисляется, исходя из количества слов в тексте и в предложении, а также слогов в словах. Полученное значение находится в диапазоне от 0 (очень низкий уровень удобочитаемости) до 100 (очень высокий) [9]. Предлагается считать, что значение индекса от 90 до 100 соответствует уровню образования пятиклассника, а, например, значение от 0 до 30 – уровню студента вуза.

Процесс идентификации потенциального адресата текста подразумевает обращение к некому набору «эталонов» – базе знаний, отражающей характерные черты текстов, предназначенных для той или иной категории читателей (как с точки зрения словаря, так и с точки зрения синтаксиса) [10]. Для текста с неизвестной категорией будет требоваться определить его наиболее вероятный класс, то есть соотнести с одним из известных классов или с несколькими из них.

4. Заключение

В работе вкратце рассмотрена проблема реализации классификации текстов по категориям реципиентов и существующие научные направления, затрагивающие решение этой задачи на основе лексических и синтаксических характеристик текста. Предложено реализовать систему классификации текстов на основе распознавания их потенциальных адресатов.

Список литературы

1. Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы. – М.: Либроком (Editorial URSS), 2009. – 264 с.

2. Sebastiani F. Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34, No. 1, pp. 1-47. – 2002.

3. Du R., RSafavi-Naini R., Susilo W. Web Filtering Using Text Classification, Proceedings of the 11th IEEE International Conference on Network (ICON 2003), pp. 325-330 – 2003.

4. Lipka N. Modeling Non-Standard Text Classification Tasks. – Weimar, Germany: Bauhaus-Universität Weimar, 2013. – 158 p.

5. Лотман Ю.М. Внутри мыслящих миров. – С.-Петербург: Языки русской культуры, 2000. – 464 с.

6. Каменская О.Л. Текст и коммуникация. – М.: Высшая школа, 1990. – 78 с.

7. Stephens C. All About Readability: [Электронный ресурс] // Plain Language. 2007-2010. URL: http://plainlanguage.com/newreadability.html (Дата обраще-ния: 19.03.2012).

8. DuBay W. The Principles of Readability: [Электронный ресурс] // Plain Language at Work Newsletter. 2013-2014. URL: http://www.impact-information.com/impactinfo/readability02.pdf (Дата обраще-ния: 19.03.2012).

9. Оборнева И. В. Автоматизация оценки качества восприятия текста. Вестник Московского городского педагогического университета, 2(5). – 2005.

10. Глазкова А.В. Возможность автоматического определения адресата на основе семантико-синтаксических особенностей текста / Открытые семантиче-ские технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2014): материалы IV междунар. науч.-техн. конф. – Минск: БГУИР, 2014. – 576 с.

ПРОЦЕДУРА ФОРМИРОВАНИЯ ИНВАРИАНТНОЙ РЕПРЕЗЕНТАЦИИ СИТУАЦИИ ДЛЯ АВТОМАТИЗАЦИИ ПРОЦЕССА ПОНИМАНИЯ ТЕКСТА В СИСТЕМЕ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ

Лазаренко О.В.
Харьковский гуманитарный университет «Народная украинская академия»
г. Харьков, ул. Лермонтовская, 27, тел. 050-300-82-83
e-mail: lazolvlad@yandex.ru

Работа над созданием искусственного интеллекта (ИИ) длится не первое десятилетие. Однако ключевая задача - научить компьютеры работать подобно человеческому мозгу - так же далека от своего решения, как и раньше. Одна из причин такого положения состоит в том, что разработчики ИИ, по мнению известного разработчика компьютеров в Силиконовой долине Джеффа Хокинса, «хотят достичь поставленной цели, обойдя вниманием вопрос о сути разума, о том, что означает слово “понимать”... Этим самым они “выплеснули с водой ребенка” — создавая мыслящие механизмы, забыли о разуме! Но все попытки создания искусственного интеллекта без учета особенностей естественного обречены на провал» [1].

Сравнивая работу человеческого мозга с работой компьютера, Дж. Хокинс задался вопросом - *какая составляющая разума отсутствует в компьютере?*

В поисках ответа на этот вопрос в августе 2002 года он открыл научно-исследовательский центр по изучению мозга, в котором первостепенное значение уделили изучению неокортекса — части головного мозга человека, ответственной за интеллект [1].

В ходе изучения неокортекса были обнаружены определенные особенности его работы, среди которых для нас представляет особый интерес способность мозга использовать инвариантные репрезентации объекта с сохранением его наиболее важных признаков на основе относительных измерений, пропорций и других характеристик, в которых возможны существенные упущения в сравнении с конкретным образом. Оказалось, что мозг запоминает важные взаимосвязи внешнего мира, а не привязывается к отдельным его элементам.

Аналогичным образом действует человеческий мозг и при чтении текстовой информации. Согласно гипотезе, выдвинутой голландским ученым А. ван Дейком о том, что при чтении текста люди часто обрабатывают информацию не полностью или неточно и, тем не менее, понимают текст. «Языковому пользователю нет необходимости дожидаться конца абзаца, главы или целого текста, чтобы понять, о чем идет речь в тексте или в его фрагменте, … пользователь языка может догадаться о теме текста уже после минимума текстовой информации из первых пропозиций. Догадку может подтвердить самая различная информация: заглавие, тематические слова, тематические первые предложения…» [2], то есть наиболее важные аспекты и их взаимосвязи.

В своих исследованиях процесса понимания текста [3] мы пришли к необходимости разработки ситуационных моделей, позволяющих уйти от использования онтологий и упростить автоматизацию процесса понимания текстов.

В разрабатываемой нами системе автоматического реферирования ситуационная модель формируется в виде накопителя текстовых баз определенной тематики, автоматически извлекаемых из текста в процессе его смыслового анализа в соответствии с разработанным алгоритмом извлечения основных смысловых аспектов текста.

Полученная таким образом ситуационная модель является основой для создания *инвариантной репрезентации ситуации*, представляющей собой набор наиболее важных признаков, выделенных на основе относительных характеристик ситуации, в которых возможны существенные упущения в сравнении с конкретной ситуацией, описываемой в конкретном тексте.

Возникает вопрос: какие признаки считать наиболее важными для описания данной ситуации, чтобы включить их в инвариантную репрезентацию?

Существенную помощь в этом могут оказать заголовки статей как концептуальные инварианты текстов, на базе которых создавалась данная ситуационная модель.

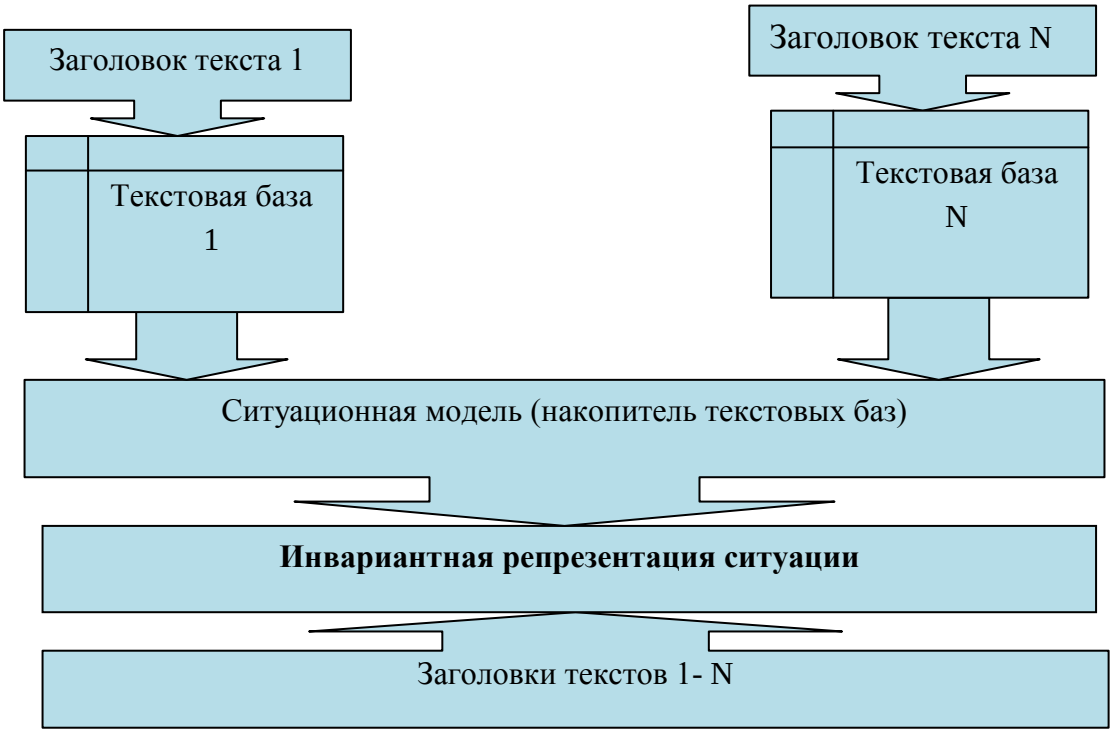


Рис. 1. Процедура разработки инвариантной репрезентации ситуации в системе автоматического реферирования

Таким образом, с использованием текстовых баз, задающих контекстную семантику, и формируемых на их основе ситуационных моделей, содержащих информацию, актуализируемую в процессе понимания текста, а также заголовков всех текстов можно выделить наиболее важные смысловые составляющие определенной ситуации, которые и составят инвариантную репрезентацию ситуации.

Может показаться, что все это уводит нас от основной задачи исследований. Однако это не так. Чем глубже и яснее мы представляем процесс понимания, осуществляемый человеком, тем точнее будет настройка системы автоматического реферирования на анализ смысла исходного текста. Помимо этого исследование и моделирование некоторых аспектов процесса понимания приоткрывают нам тайны работы мозга. Мы видим, как работы в разных областях, таких как исследование механизмов работы мозга (Дж. Хокинс), разработка стратегий понимания дискурса (А. ван Дейк) и моделирование процесса реферирования (Лазаренко О.В.) сошлись в одной точке – инвариантных представлениях, лежащих в основе указанных процессов. Результаты этих исследований подтверждают прямо и косвенно тот факт, что мозг в процессе распознавания объекта, фактов, ситуаций и проч. вспоминает важные взаимосвязи внешнего мира, а не привязывается к отдельным его элементам.

Все вышеизложенное позволяет надеяться, что разрабатываемая нами процедура смыслового анализа текста позволит обеспечить более качественный результат автоматического реферирования за счет использования доступа к информации необходимой для синтеза реферата путем:

1. выделения макроструктуры текста и формирования на ее основе текстовой базы;
2. формирования в автоматическом режиме ситуационных моделей в виде накопителей текстовых баз, используемых для более точного понимания смысла конкретного текста;
3. извлечения знания, имеющегося в тексте, через актуализацию его с помощью инвариантных репрезентаций ситуаций.

Список литературы

1. *Хокинс Дж.*, Блейксли С. Об интеллекте / Дж., Хокинс, Блейксли С. - М.: Издательский дом "Вильямс", 2007. - 240 с.
2. *Дейк ван Т. А.* Стратегии понимания связного текста / Т. А. ван Дейк, В. Кинч // Новое в зарубежной лингвистике. – Вып. 23: Когнитивные аспекты языка. – М., 1988. – С. 153–211.
3. *Лазаренко О.В.* Разработка интеллектуальной системы автоматического реферирования с использованием текстовых баз и ситуационных моделей /О.В. Лазаренко // MegaLing’2013. Горизонти прикладної лінгвістики та лінгвистичних технологій : доп. міжнар. наук. конф., Україна, Київ, 20-23 листопаду 2013 г.

МОДЕЛЬ БАЗЫ ЗНАНИЙ АВТОМАТИЗИРОВАННОЙ ИНФОРМАЦИОННОЙ БИБЛИОТЕЧНОЙ СИСТЕМЫ

Кочуева З.А.

Национальный технический университет
"Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: *kochueva@kochuev.com*

Современные библиотеки являются знание-ориентированными информационными системами, оперирующими естественно-языковыми объектами. Автоматизация процессов по обработке данных в библиотеках обусловлена непрерывно увеличивающимся объемом различной информации во всех отраслях человеческой деятельности. Это обуславливает актуальность развития информационных технологий, моделей, методов, алгоритмов и программ интеллектуальной обработки данных и применение этих методов для автоматизированной обработки языковой информации в библиотечных системах.

Предложена модель базы знаний АИБС на основе построения семантической сети понятий, записанных на языке алгебры конечных предикатов. Модель предметной области содержит концепты понятий, объектов и отношений из области компьютерных технологий и Интернет. Рассматриваемая область знаний довольно хорошо структурирована, поэтому ее можно адекватно отобразить иерархической семантической сетью (ИСС). Модель, основанная на использовании ИСС, опирается на принципы организации человеческой памяти. Под семантической сетью понимают систему знаний, представленную в виде целостного образа сети, узлы которой отвечают понятиям и объектам, а дуги – отношениям между объектами. Используем ИСС как способ представления семантических отношений между концептами. На первом шаге для создания иерархической системы знаний выделяем основные объекты предметной области. Их в нашей модели - 34.

Понятия в данной модели реализуют отношения трех типов: "быть частью" (PART OF); "является" (IS-A), "имеет" (HAS). Для отображения иерархических отношений между точками соприкосновения концептов, а также для установления связей между узлами, показывающими концепты и их экземпляры, используются отношения IS-A. Отношение IS-A передает наследование атрибутов между уровнями иерархии, т.е. отношение IS-A является отношением включения или совпадения понятий.

Универсум U нашей задачи представлен множеством всех понятий, входящих в базу знаний системы. При переходе от графического изображения к конечным предикатам введены два бинарных предиката: предикат $P(x_1,x_2)$ - описывает отношения вида «часть – целое»: он равняется 1, если x_1,x_2 находятся в отношении «часть – целое», и равняется 0 в противоположном случае. Предикат $Q(x_1,x_2)$ – описывает отношения вида «является» (IS-A): он равняется 1, если

x_1,x_2 находятся в отношении «род – вид», и 0 в противоположном случае. Пример графического представления этих отношений представлен на рисунке 1.

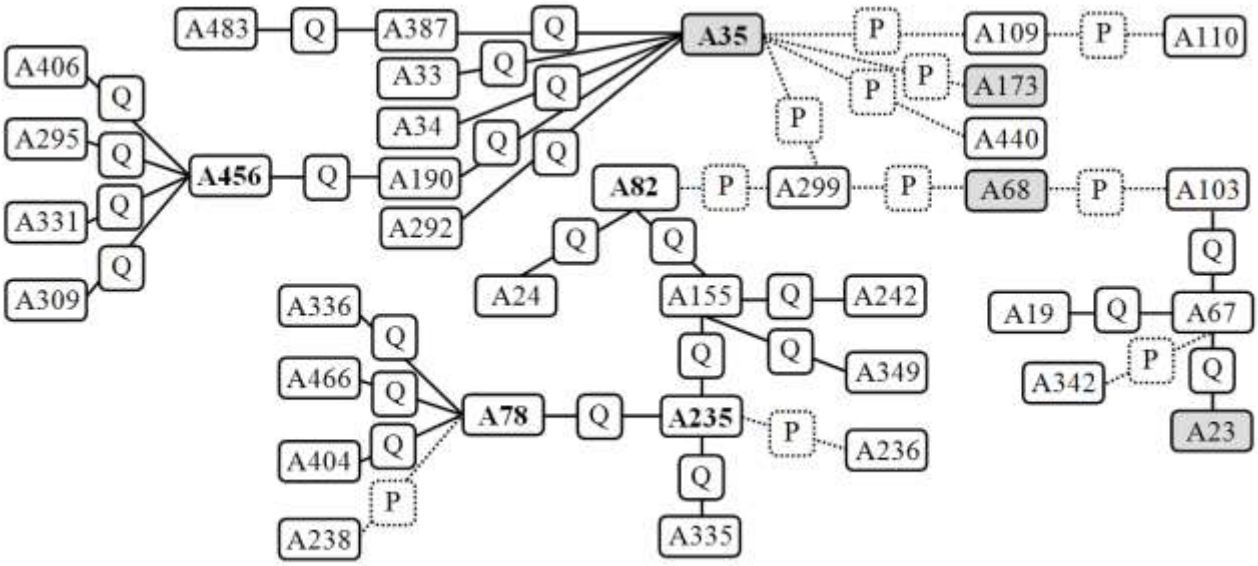


Рисунок 3 – Графическое представление отношений между понятиями предметной области

Каждая пара вершин сети упорядочена одним из отношений, соответствующих применяемым в модели предикатам, каждому предикату соответствует дуга графа, соединяющая вершины, в которых помещаются его переменные (или их значения). В предикативной записи $P(x_1,x_2)$ и $Q(x_1,x_2)$, два элемента x_1, x_2 находятся в некотором отношении зависимости, когда один является основным (x_1), другой – подчиненным (x_2).

Такая модель предметной области позволяет получить семантическое представления текста, относящегося к данной предметной области.

Предложенные программные средства, предотвращают увеличение объема базы данных, позволяют ускорить обработку полнотекстовых документов в библиотеке, сократить время ожидания документа в процессе его обработки.

Список літератури

1. . *Кочуева, З. А.* Моделирование процедур систематизации и классификации информационных объектов методом компараторной идентификации [Текст] / Н. В. Борисова, З. А. Кочуева, Н. В. Шаронова, Н.Ф. Хайрова // Вестник Херсонского национального технического университета. – Херсон : ХНТУ. – 2012. – № 1(44). – С. 91-95

2. *Кочуева З. А.* Индексирование полнотекстовых документов для задачи интеллектуального поиска информации по ключевым словам [Текст] / З. А. Кочуева, Н. В. Борисова // Східно-Європейський журнал передових технологій. – Харків : ПП «Технологічний Центр», 2014. – № 1/2 (67). – С. 4-8.

3. *Хайрова Н. Ф.* Автоматизированные информационные системы: задачи обработки информации [Текст] / Н. Ф. Хайрова, Н. В. Шаронова – Х.: ХГУ «НУА», 2002. – 120 с.

ИНТЕЛЛЕКТУАЛЬНЫЕ ТЕХНОЛОГИИ ИДЕНТИФИКАЦИИ ФАКТОГРАФИЧЕСКОЙ ИНФОРМАЦИИ

Дорошенко А.Ю.
Национальный технический университет НТУ «ХПИ»
г. Харьков, ул. Пушкинская, 79/2, тел. 707-63-60,
e-mail: doroshenkoanastasiia@gmail.com

В работе обсуждается технология фактографического поиска, предлагается подход, основанный на представлении содержания текста в форме семантической сети, позволяющий искать факт в семантической сети определенного текста. Рассмотрена задача экстракции и идентификации знаний фактографической информации об ЭВМ.

Цель работы – построение семантической сети фактографической информации об ЭВМ на основе анализа и обработки текстов, а также решение задачи с помощью алгебры предикатов.

В общепринятом смысле под семантической сетью понимается модель представления знаний посредством сети узлов, связанных дугами, где узлы соответствуют понятиям или объектам, а дуги – отношениям между узлами [5].

Научной основой построения семантических сетей является теория графов. Семантические сети представляют знания в виде графовой структуры, которая является более наглядной и естественной по сравнению с другими структурами знаний. Решаемая задача использует структуру, моделирующую семантические связи, которые мы используем для получения одних фактов на основе других [5]. Построение графа помогает находить противоречия в знаниях, а также выявлять недостающие фрагменты знаний. Среди особенностей семантических сетей можем выделить:

- описание объектов ПрО (полной семантической сети) осуществляется средствами естественного языка;
- все факты, включая и вновь поступившие, накапливаются в относительно однородной структуре памяти;
- на сетях определяют ряд унифицированных семантических отношений между объектами и соответственно унифицированные методы вывода;
- структурное представление семантических знаний позволяет определить на них дополнительную семантику, определяющую относительную силу семантических связей, облегчающую процесс вывода в сетях.

К фактографической информации мы относим информацию о фактах [3]. Фактографическую информацию обычно сознательно трактуют просто как конкретные сведения или данные независимо от того, являются ли они фактическими или прогнозируемыми. Главное, что эти сведения сообщают о какой-то предметной области, а не о документах, посвященных этой области. Исходя из такого понимания, фактографическую информацию можно классифицировать следующим образом:

1) фактическая и прогнозная (гипотетическая) информация;

2) количественная и качественная фактографическая информация;

3) хорошо структурированная фактографическая информация и плохо (слабо) структурированная фактографическая информация.

К хорошо структурированным сведениям об ЭВМ относятся, прежде всего, сведения количественного характера, а также качественные (словесно выраженные) сведения, имеющие хорошо регламентированную форму: параметры оборудования и их значения и т. п. К плохо структурированным относятся сведения, представленные разнообразными нерегламентированными словесными инструкциями, т. е. различные описания отдельных фактов, изложение концепций и теорий, сделанных на естественном языке [2,3].

Заключение. Знание не такое определенное понятие, как факт. Оно лишь ограничивает множество возможных состояний мест предметного пространства. Поиск факта есть поиск в семантической сети текста такой подсети, которая изоморфна одному из шаблонов. Если подсеть найдена, факт считается установленным, после чего производится извлечение сущностей и их маркировка ролями, заданными в соответствующих узлах лингвистических описаний [1,4]. Таким образом, результатом поиска является имя факта и набор указателей на сущности семантической сети с указанием соответствующих им ролей в лингвистическом описании.

Список литературы

1. *Алисейко З. А.* Использование алгебры предикатов и предикатных операций для формализации декларативной и процедурной составляющих знаний / З. А. Алисейко, В. И. Булкин, О. В. Канищева, Н. В. Шаронова // Біоніка інтелекту. – Харків : ХНУРЕ, 2006. – № 1(64). – С. 59-63.

2. *Амамия М.* Архитектура ЭВМ и искусственный интеллект / М. Амамия, Ю. Танака. – М. : Мир, 1993. – 400 с.

3. *Бондаренко М. Ф.* О мозгоподобных ЭВМ / М. Ф. Бондаренко, З.В. Дударь, И.А. Ефимова, В.А. Лещинский, С.Ю. Шабанов-Кушнаренко // Радиоэлектроника и информатика. – Харьков : ХНУРЭ, 2004. – № 2. – С. 89-105.

4. *Булкин В.И.* Математические модели знаний и их реализация с помощью алгебропредикатных структур / В. И. Булкин, Н.В. Шаронова: монография. – НТУ «ХПИ», МЭГИ. : Донецк, 2010. – 304 с.

5. *Sussna M.* Word sense disambiguation for free-text indexing using a massive semantic network // The Second International Conference on Information and Knowledge Management. – 1993. – P. 67-74.

НА ШЛЯХУ ДО СТВОРЕННЯ АНГЛІЙСЬКО-УКРАЇНСЬКОГО ЕЛЕКТРОННОГО СЛОВНИКА З ЕНЕРГЕТИЧНОГО МАШИНОБУДУВАННЯ

Купріянов Є. В.
Національний технічний університет
«Харківський політехнічний інститут»,
м. Харків, вул. Пушкінська 79/2, тел. 707-63-60,
e-mail: cuprijanow.eugen@yandex.ua

Сьогодні Україна розширює співробітництво з іноземними країнами у виробничих галузях, зокрема в енергетичному машинобудуванні. У цьому зв'язку обсяг інформації, яку необхідно опрацьовувати, постійно зростає. Коло завдань її опрацювання також охоплює переклад технічної документації, для виконання якого застосовують електронні словники спеціальної лексики. Точний та оперативний переклад залежить не лише від кваліфікації та досвіду самого перекладача, але й також від якісно укладеного словника.

Наявні електронні вузькогалузеві словники, зокрема «ABBYY Lingvo», «Мултитран», «Мултилекс» та інші не можуть задовольнити усі потреби користувача, а саме: 1) семантичні особливості технічного терміна залежно від його предметної співвіднесеності; 2) особливості функціонування терміна у контексті, його здатність утворювати оказіональні відповідники в мові перекладу; 3) енциклопедична інформація (текстова, графічна й анімаційна) про предмети та процеси позначувані терміном. Отже, **актуальною** постає проблема вироблення нового підходу до створення електронних лексикографічних праць, що враховують зазначені потреби під час перекладу технічних текстів.

Мета нашої розвідки – запропонувати новий підхід до створення електронних словників спеціальної лексики та реалізувати його у вигляді програмного продукту. Для досягнення мети необхідно: 1) розглянути головні етапи технічного перекладу та визначити інформаційні потреби перекладача на цих етапах; 2) встановити семантичні характеристики енергомашинобудівних термінів (синонімія, омонімія, полісемія), а також можливість утворення видових номінацій і сталих термінологічних сполучень залежно від предметної сфери; 3) вибрати контексти, що наочно демонструють особливості вживання термінів досліджуваної галузі; 4) розробити структуру словникової статті та обґрунтувати вибір лексикографічних параметрів опису терміна, і 5) створити програмну оболонку словника. **Предметом** розвідки є електронне лексикографічне упорядкування термінології енергетичного машинобудування з урахуванням потреб користувача на всіх етапах перекладу.

Наукова новизна нашого дослідження полягає в тому, що розроблюваний електронний словник, на відміну від інших аналогів, призначено не лише для подання перекладних еквівалентів до англійських термінів, а й також для детального опису семантики самих англійських термінів:

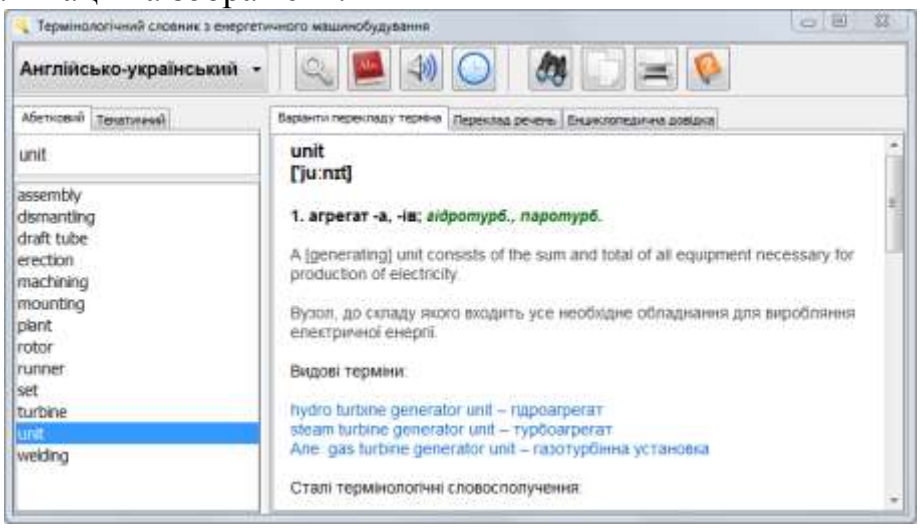
– значення заголовкового терміна в англійській мові, залежно від галузі його використання (подано англійську дефініцію із перекладом українською);

– абсолютні або часткові синоніми;

– наявність видових номінацій і сталих термінологічних сполучень, утворюваних англійським терміном у певному значенні (супроводжуються перекладом українською мовою);

– здатність утворювати полісемічні та омонімічні зв'язки з термінами як у межах однієї, так і кількох термінологій неспоріднених галузей (подано відповідники).

Крім опису семантичних характеристик передбачена також демонстрація функціонування англійських термінів у різних контекстах у певному значенні. Для цього передбачено кольорове виокремлення описуваного терміна та його контекстного оточення. Правильний переклад потребує від перекладача фонових знань. Саме тому електронний словник містить енциклопедичну інформацію як додатковий засіб семантизації. Довідкова інформація може бути у вигляді тексту, анімації та зображень.



Зовнішній вигляд електронного словника

Висновки. Розроблюваний електронний словник можливо використовувати на всіх етапах перекладацького процесу: 1) аналіз тексту, написаного мовою оригіналу, з метою його перекодування; 2) перекодування, тобто підстановка знаків мови перекладу замість знаків мови оригіналу; 3) реалізація тексту мовою перекладу.

Список літератури

1. *Гарбовский Н. К.* Теория перевода: [Учебник] / Н. К. Гарбовский. – М.: Изд-во Моск. ун-та, 2007. – 544 с.

2. *Герд А. С.* Основы научно-технической лексикографии / А. С. Герд. – Ленинград : Изд-во Ленинград. ун-та, 1986. – 72 с.

3. *Гринев-Гриневич С. В.* Введение в терминографию: Как просто и легко составить словарь : [Учебное пособие] / С. В. Гринев-Гриневич. – М.,2009. – 224 с.

4. *Карабан В.* Переклад англійської наукової і технічної літератури. Граматичні труднощі, лексичні, термінологічні та жанрово-стилістичні проблеми / В. Карабан. – Вінниця : Нова Книга, 2004. – 576 с.

5. *Цвиллинг М. Я.* О некоторых вопросах технической лексикографии // Тетради переводчика. – М., 1976. – № 13., – С. 115–127.

УКРАЇНСЬКИЙ СКЛАДОПОДІЛ У СВІТЛІ СЕГМЕНТАЦІЇ МЕРМЕЛЬШТАЙНА (ЕКСПЕРИМЕНТАЛЬНО-ФОНЕТИЧНЕ ДОСЛІДЖЕННЯ)

Іщенко О.С.
Інститут української мови НАН України
Київ, вул. Грушевського, 4, тел. 0 (44) 278-18-85,
e-mail: o.ishenko@gmail.com

Згідно зі спостереженнями П. Мермельштайна [5] поділ потоку мовлення на склади відбувається у місцях суттєвого спаду інтенсивності звукової хвилі між сегментами, тривалість яких є типовою для цих одиниць (складів). Ядром же складу є точка максимального рівня інтенсивності в межах сегментів. Таке розуміння складу є суто акустичним. Щоб поділити мовлення на склади за цим принципом, необхідно послуговуватися інструментами, які традиційно використовуються у фонетичних дослідженнях (ідеться про програмне забезпечення для аналізу звукових коливань), оскільки засобами слуху відчути всі зміни інтенсивності звукової хвилі не можливо.

Сегментація мовлення на склади за правилом Мермельштайна має широке практичне застосування. Так, у низці праць[1], у яких була необхідність ділити мовлення на склади, використано сегментацію саме за цим правилом. Підкреслимо, що П. Мермельштайн винайшов спосіб сегментації мовлення на склади, вивчаючи англійську мову. Дослідження, в яких застосовано даний алгоритм, здійснені на прикладі неслов'янських мов.

Мета нашого експериментального дослідження – перевірити правило Мермельштайна на матеріалі української мови[2] та підтвердити чи спростувати можливість його застосування у фонетичних студіях з лінгвоукраїністики. Для цього залучено зразки[3] літературного й діалектного мовлення.

Дослідження побудовано на основі зіставлення контрольної та експериментальної груп зі зразками сегментованого мовлення. У контрольній групі сегментацію мовлення виконано згідно з традиційними принципами українського складоподілу [2]; в експериментальній групі[4] – за принципами алгоритму Мермельштайна.

Результати дослідження засвідчили доволі сильну розбіжність складової сегментації мовлення в контрольній та експериментальній групах. Математично розбіжність становить 40%.

Серед основних розбіжностей зафіксовано:

- чітку акустичну сегментацію глухих шумних звуків. Див. склади [че] і [ка] на рис. 1. Себто, згідно з правилами П. Мермельштайна, такі

[1] Див. праці: [4], [6], [7] та ін.
[2] Мета дослідження зумовлена розумінням того, що складотворення в кожній мові має індивідуальну природу.
[3] Використано аудіозаписи усного мовлення загальною тривалістю 60 хв. Джерела: [1], [3].
[4] Сегментацію мовлення на склади за принципами алгоритму Мермельштайна здійснено в ПЗ Speech Analyzer [8].

приголосні, як [ч] і [к], варто вважати складовими (складотворчими), що суперечить фонетичному розумінню цих звуків;

- спорадичну (20% випадків) сегментацію нескладових [ў] та [ĭ], зокрема в позиції після паузи (на рис. 1 бачимо сегментацію [ў] за зміною інтенсивності коливань). Тобто, з погляду концепції П. Мермельштайна, так звані нескладові звуки можна вважати складовими (принаймні, позиційно);
- регулярну (75% випадків) відсутність сегментації між двома сусідніми голосними (див. однофонемний склад [о] на рис. 1);
- регулярну (70% випадків) відсутність сегментації між голосним і наступною послідовністю шумного й сонорного, що вимагається правилами українського складоподілу [2, 352]. Так, на рис. 1 нормативний складоподіл [ві’-дм’ін] не відповідає акустичному [від-м’ін], оскільки звук [д] перебуває в зоні спадання інтенсивності звукових коливань, розпочатої в межах голосного [і]. Це саме стосується сегментації слова [ору́дного], де [д] і [н] належать різним акустичним сегментам;
- регулярну відсутність сегментації між двома сонорними. Зокрема, у 60% випадків два сусідні сонорні реалізовані фактично без зміни інтенсивності акустичних коливань, що свідчить про належність їх до одного спільного сегменту (згідно з нормою вони належать різним складам [2, 353]).

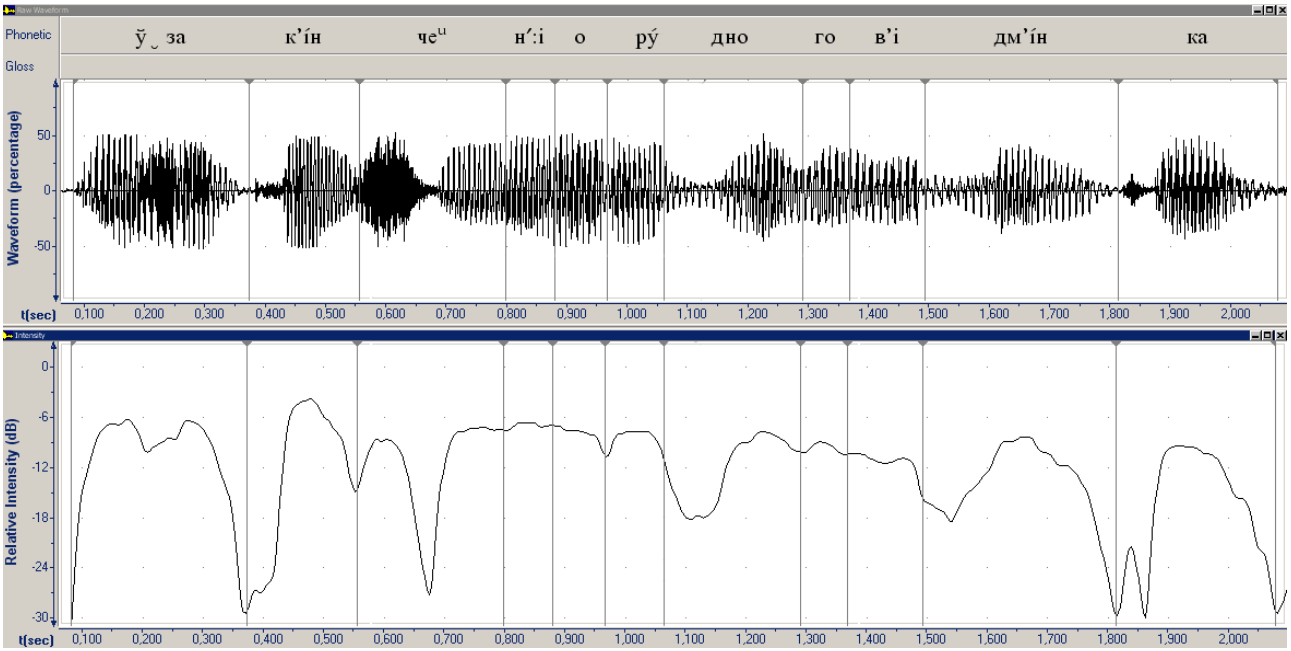


Рис. 1: Складоподіл синтагми [ў◡зак’інчеᵘн′:і ору́дного в’ідм’інка] на тлі її акустичної експлікації

Загалом кореляція традиційного розуміння складоподілу та акустичної сегментації П. Мермельштайна найтісніша тоді, коли зіставляємо відкриті склади типу CV[5].

Отже, розглянутий алгоритм членування мовлення на склади не може бути застосовано у чистому вигляді для української мови (ґрунтуючись лише на аналізі зміни рівня інтенсивності звукових коливань). Водночас представлене акустичне розуміння складу дає можливість по-новому оцінити суперечливі правила складоподілу, пов'язані з нескладовими звуками, поєднанням шумних і сонорних звуків, сонорних і сонорних тощо, оскільки чинні норми складоподілу були сформовані і з урахуванням акустичного аспекту мовлення.

Список літератури

1. Погрібний М. Українська літературна вимова. Дніпропетровськ: Трансформ, 1992. – 26 с. [з аудіоносієм].

2. Сучасна українська літературна мова. Фонетика / Відп. ред. М.А. Жовтобрюх. – К.: Наук. думка, 1969. – 436 с.

3. Український діалектний фонофонд / П.Ю. Гриценко та ін. – К.: Ін-т української мови НАН України, 2004. – 168 с.

4. Jong N. Praat script to detect syllable nuclei and measure speech rate automatically / N. Jong, T. Wempe // Behavior Research Methods. – 2009. – Vol. 41 (2). – P. 385–390.

5. Mermelstein P. Automatic segmentation of speech into syllabic units / P. Mermelstein // Journal of the Acoustical Society of America. – 1975. – Vol. 58. – P. 880–883.

6. Pfau T. Estimating the speaking rate by vowel detection / T. Pfau, G. Ruske // Acoustics, Speech, and Signal Processing: Proceedings of the IEEE'1998). – 1998. – Vol. 2. – P. 945–948.

7. Pfitzinger H. Local speech rate perception in German speech / H. Pfitzinger // Proceedings of the XIVth International Congress of Phonetic Sciences. – 1999. – Vol. 2. – P. 893–896.

8. Speech Analyzer: computer program for acoustic analysis of speech sounds // SIL International Inc. – Version 3.1. – 2012. – < http://www-01.sil.org/computIng/sa >.

[5] CV – схема складу "приголосний + голосний"; традиційний (найбільш поширений) тип складу в українській мові.

ЗАДАЧА ДОСЛІДЖЕННЯ ВПЛИВУ КОНТЕКСТУ НА ВИБІР ЛЕКСИЧНИХ ПАРАЛЕЛЕЙ ПРИ АВТОМАТИЗОВАНОМУ ПЕРЕКЛАДІ

Цибізова Ю. С.
Національний технічний університет
«Харківський політехнічний інститут»
м. Харків, вул. Фрунзе, 21, тел. (057) 700-15-64,
e-mail: omsroot@kpi.kharkov.ua

Виступ присвячено короткому аналізу словників лексичних паралелей, теоретичним положенням вчених, які досліджували тему зовні схожих слів у різних мовах, які можуть повністю / частково / неповністю збігатися за значенням та контексту при перекладі таких багатозначних слів.

Першим словником, який, був проаналізований, був «Англо -російський і російсько-англійський словник" хибних друзів перекладача"» В.В.Акуленко (М., 1969) [1]. **Міжмовні синоніми** – це слова двох мов, які повністю або частково збігаються за значенням і вживанням (і, відповідно, які є еквівалентами при перекладі). **Міжмовними омонімами** можна назвати слова двох мов, подібні до ступеня ототожнення за звуковою (або графічною) формою, але які мають різні типи значення. Нарешті, до **міжмовних паронімів** слід віднести слова порівнюваних мов, не цілком подібні за формою, але які можуть викликати хибні асоціації та ототожнюватися один з одним, незважаючи на фактичне розходження їх значень.

Другий словник, який описує дану лексику, - «Німецько - російський і російсько - німецький словник "хибних друзів перекладача "» К.Г.М.Готліба (М., 1972) [2]. Розглядаючи **міжмовні аналогізми** з точки зору їх зовнішньої структури, автор визначає ступінь їх фоно - морфологічної, графічної та семантичної близькості.

Цю ж проблему досліджував проф. М.П.Кочерган , створюючи «Словник російсько - українських міжмовних омонімів» (К. , 1997) [3]. **Міжмовні омоніми** – це слова двох контактуючих мов, які повністю чи здебільшого збігаються за формою, але розрізняються за змістом. Наприклад: рус. *неделя* «сім днів , тиждень», укр . *неділя* «останній день тижня».

Проаналізувавши всі ці словники, було вирішено, що найкраще зупинитися на терміні "лексичні паралелі", який був запропонований в 1993 р. проф . В.В.Дубічинським [4]. Лексеми, що збігаються в плані вираження і подібні / несхожі в плані змісту називаються узагальнюючим терміном "**лексичні паралелі**". Якщо зовні подібні лексеми порівнюваних мов семантично повністю збігаються, то такі лексичні паралелі називаються **повними**. Наприклад, укр. *архітектура* і нім. *Architectur*. У разі збігу одних і неспівпадіння інших значень семантичних структур зовні схожих лексем, мова йде про **неповні лексичні паралелі**. Наприклад, укр. *диктат* і нім. *Diktat*.

У даній класифікації значення, які збігаються прийнято називати **інтерсememами**, а ті, які не співпадають та відображають національно -

культурну своєрідність лексичної одиниці – **ідиосемемами.** Поняття інтерсемем і ідиосемем дає можливість провести порівняльний (перекладацький) аналіз на рівні окремих значень лексем і в певних випадках врахувати навіть семантичні та стилістичні нюанси на рівні дрібніших компонентів значень – сем. У разі ж розбіжності всіх значень зовні схожих лексичних одиниць двох або більше порівнюваних мов, йдеться про **хибні лексичні паралелі.** Наприклад, укр. *актор* та нім. *Akteur*.

При перекладі текстів труднощі виникають здебільшого саме з багатозначними словами. Який саме варіант обрати залежить від багатьох факторів, але найчастіше посилання роблять саме на контекст, відзначаючи, що вибір тієї чи іншої відповідності при перекладі багато в чому визначається контекстом, в якому вжита та чи інша мовна одиниця.

Під контекстом прийнято розуміти мовне оточення, в якому вживається та чи інша лінгвістична одиниця.

Контекст – закінчений уривок письмової або усної мови (тексту), загальний зміст якого дозволяє уточнити значення окремих слів, речень, які входять до нього.

У межах загального поняття контексту розрізняється вузький контекст (або «мікроконтекст») і широкий контекст (або «макроконтекст»). Під **вузьким контекстом** мається на увазі контекст речення, тобто лінгвістичні одиниці, складові оточення даної одиниці в межах речення. Під **широким контекстом** мається на увазі мовне оточення даної одиниці, що виходить за рамки речення; це – текстовий контекст, тобто сукупність мовних одиниць, оточуючих цю одиницю в межах, що лежать поза даним реченням, іншими словами, в суміжних з ним реченнями.

Так, контекстом слова є сукупність слів, граматичних форм і конструкцій, в оточенні яких зустрічається дане слово. Контекст є одним з головних чинників якісного перекладу лексичних паралелей.

Список літератури

1. *Акуленко В.В.* Англо-русский и русско-английский словарь “ложных друзей переводчика” - М., 1969.

2. *Готлиб К.Г.М.* Немецко-русский и русско-немецкий словарь “ложных друзей переводчика” - М., 1972.

3. *Кочерган М.П.* Словник російсько- українських міжмовних омонімів - К., 1997.

4. *Дубичинский В.В.* Лексические параллели – Харьков, 1993; *Дубичинский В.В., Ройтер Т.* Русско-немецкий словарь лексических параллелей – М., 2011.

5. *Одинцов В.В. Стилістика тексту. М., 1980.*

6. *Ахманова О.С., Гюббенет І.В.* Вертикальний контекст як філологічна проблема / / Питання мовознавства , № 3 , 1977.

АВТОМАТИЗОВАНЕ ВИДОБУВАННЯ ТЕРМІНОЛОГІЧНИХ ОДИНИЦЬ З НАУКОВО-ТЕХНІЧНИХ ТЕКСТІВ

Борисова Н. В., Решетило С. С.
Національний технічний університет
"Харківський політехнічний інститут",
м. Харків, вул. Пушкінська, 79/2, тел. 707–63–60
e-mail: *borisova_nv@mail.ru, 13-svetik.cat.07@mail.ru*

Вирішення багатьох задач автоматизованої обробки текстів потребує видобування з текстів термінів, тобто слів або словосполучень, що називають поняття певної предметної області. Науковий термін не тільки точно і однозначно визначає чітко окреслене спеціальне поняття будь-якої галузі науки, а й відображає його співвідношення з іншими поняттями в межах предметної області.

Видобування термінів необхідне при вирішенні багатьох задач автоматизованої обробки текстів, а саме машинний переклад, літературно-наукове редагування, видобування знань з наукових текстів, реферування та анотування текстів, складання словників певної предметної області та ін.

Оскільки наявність термінів, понять та їх визначень є особливістю науково-технічних текстів, тому що основною функцією наукового стилю є оформлення, збереження, передача наукової інформації, для виявлення ознак термінів доцільно було б проаналізувати лексико-фразеологічні та дискурсивні особливості цих текстів [1]. Це необхідно для виділення дискурсивних маркерів, які відповідають дискурсійній операції «визначення понять». Саме ці маркери і є ознаками термінів у науково-технічних текстах (табл. 1).

Таблиця 1 – Приклади використання деяких груп дискурсивних маркерів

Опис	**Приклади використання**
1	2
Група «*називатися*»	
дієслово «називатися» у формі третьої особи однини теперішнього часу	***Атмосферою*** *називається зовнішня газова оболонка Землі, що сягає від її поверхні в космічний простір приблизно на 3000 км…* [2]
Група «*бути + називати*»	
дієслово "бути" в формі першої особи множини майбутнього часу	*Угрупованням тут будемо називати досить чітко окреслений та територіально сильно обмежений рівень живого* [2].
Група «*так + званий*»	
дієприкметник «званий» в різних формах	*Небажаним є досягнення так званого «екологічного імперативу» – своєрідної межи або рівня взаємодії суспільства та природи, перевищення якого буде мати катастрофічні наслідки для людства* [2]

Продовження таблиці 1

1	2
Група «*розуміти*»	
дієслово «розуміти» у формі першої особи множини теперішнього часу	*Під біоценозом екологи розуміють історично сформовану сукупність рослин, тварин та мікроорганізмів, що населяє біотоп* [2].
Група «це»	
наявність частки «це»	*Ланцюги живлення – це ряди взаємопов'язаних видів, в яких кожний попередній є об'єктом живлення наступного* [2]
Група «–»	
наявність «–»	*Біоконверсія – біологічна переробка органічних відходів промисловості, сільського й комунального господарства* [2]

Система автоматизованого видобування термінів з науково-технічних текстів певної предметної області виявлятиме терміни саме за дискурсивними маркерами. Для того щоб розробити таку систему можна використати *регулярні вирази* – систему обробки тексту, засновану на спеціальній системі запису зразків (шаблонів, масок), що задають правила пошуку. Зараз регулярні вирази використовуються багатьма текстовими редакторами і утилітами для пошуку та зміни тексту на основі вибраних правил [3].

Алгоритм автоматизованого видобування термінів з науково-технічних текстів з використанням регулярних виразів представлено нижче:

1. Відібрати для аналізу множину текстів предметної області.
2. У текстах предметної області визначити дискурсивні маркери операції «визначення поняття».
3. Розподілити дискурсивні маркери по групах.
4. Задати регулярні вирази для пошуку дискурсивних маркерів, що відповідають дискурсивній операції «визначення поняття».
5. Для кожного тексту з множини здійснити пошук термінів, використовуючи побудовані на кроці 4 регулярні вирази.
6. Вивести список термінів.

Список літератури

1. Баева Н.В. Структурирование и извлечение знаний, представленных в научных текстах / Н.В. Баева, Е.И. Большакова, Н.Э. Васильева // Девятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2004. Труды конференции в 3-х томах. Т. 2. – М.: Физматлит, 2004. – С. 46-54

2. Кучерявий В.П. Екологія. – Львів: Світ, 2001 – 500 с.: [Електронний ресурс]. – Режим доступу: http://eduknigi.com/ekol_view.php?id=1

3. Царьков В.Б. Теория и методика построения регулярных выражений. Проблема самообразования: [Електронний ресурс]. – Режим доступу: http://lipetsk.lug.ru/projects/re/re-building-howto.html

РОЗПІЗНАВАННЯ ІНФОРМАЦІЇ РЕКЛАМНОГО ЗМІСТУ У ВХІДНИХ ЕЛЕКТРОННИХ ПОВІДОМЛЕННЯХ

Ільїнський Б. В.
Національній технічний університет
«Харківський політехнічний інститут»
м. Харків, вул. Пушкінська, 79/2, тел. 707-63-60,
E-mail: bogdan.ilyinsky@gmail.com

Останнім часом зростає використання інтернет-реклами як одного з найбільш дієвих способів залучення покупця. Незважаючи на те, що найбільшим попитом на сьогоднішній день користується просування товарів і послуг за допомогою соціальних мереж, існує шлях оповіщення споживача, який можна назвати класичним для мережі Інтернет – електронна пошта (*e-mail*).

Переваги використання *e-mail* для доставки рекламних повідомлень:

- електронна пошта є практично у всіх користувачів Мережі;
- електронна пошта являє собою *push*-технологію віщання;
- електронна пошта дає можливість персоніфікованого звернення.

Цікаве, з точки зору одержувача, повідомлення може бути поширене їм серед його колег і знайомих [3].

В Інтернеті існує безліч списків розсилки (*mailing lists, "opt-in" E-mail marketing*), які присвячені різним тематикам. Одержувачі подібних листів власноруч підписалися на розсилку, та в будь-який момент у них є право і можливість скасувати свою підписку.

Існують наступні види підписок *e-mail* розсилки:

- відкриті розсилки – доступні усім бажаючим;
- закриті – призначені для використання користувачам певного кола;
- безкоштовні – що існують за рахунок ентузіазму творців, спонсорської підтримки чи платних рекламодавців;
- платні – які дозволяють використання своїх послуг за передплату.

Оскільки список розсилки звичайно є засобом віщання для визначеної цільової групи та часто має тисячі абонентів, він є ефективним інструментом маркетингу. Ряд компаній на своїх офіційних сайтах пропонує відвідувачам підписатися на розсилку, що інформує про новини компанії та оновленнях сайту. Дана розсилка нагадує абонентам про сайт та діяльність його власника, інформуючи і стимулюючи повторні візити.

Деякі компанії, що займаються легальним бізнесом, рекламують свої товари або послуги за допомогою спаму. Спам (англ. *spam*) – розсилка комерційної та іншої реклами або інших видів повідомлень особам, які не виражають бажання їх отримувати. Привабливість такої реклами – низька вартість та (приблизно) велике охоплення потенційних клієнтів.

Слід зауважити, що повідомлення, які звично називають спамом не завжди несуть у собі маркетинговий характер. До інших видів несанкціонованих повідомлень належать: «нігерійські листи» (*scam*), «фішинг» (*phishing*), «лист щас-

тя», пропагандистські листи різного характеру, *DoS*- і *DDoS*-атаки, листи, що містять комп'ютерні віруси (*malware*) та ін. [4].

Багато досліджень базуються на класифікації пошти згідно з заданими категоріями, що в загальному випадку відповідає папкам у поштовому клієнті. Серед підходів, які були застосовані для такого типу класифікації, виділяють класифікацію за допомогою машинного навчання та видобування інформації (machine learning and IR approaches). До них належать: *MailCat, SVM (support vector machines), Re:Agent, SpamCop* [1].

До підходів, які застосовувалися для фільтрування спаму, належать метод Баєса (*Bayesian approach*) та, як і у випадку класифікації поштових повідомлень, машинне навчання. Вони демонструють достатньо високу точність при визначенні спаму. Більшість систем, побудованих за таким зразком, використовують попереднє опрацювання даних, а саме, токенізацію та стемінг (*tokenization and stemming*). На етапі токенізації визначаються токени, себто елементи, які будуть використовуватися під час навчання. Ними зазвичай є слова. Тобто токенізація передбачає усунення пунктуації й екстрагування слів. Оскільки одне і те ж слово може використовуватися у різних формах (різних відмінках, різному числі тощо), необхідно здійснити стемінг, тобто виділити корінь/основу слова або представити слова лише у начальній формі: інфінітиві (для дієслів), називному відмінкові однини (для іменників, займенників) тощо. Майже в усіх дослідженнях використовувалося вже розроблене програмне забезпечення, що здійснює стемінг [2].

Результатом даного дослідження стало програмне забезпечення, що автоматично розподіляє вхідну кореспонденцію користувача по каталогах. Принципи роботи програми базуються на структурних властивостях маркетингових та інших видів повідомлень (усього розглянуто приблизно 1,5 тисячі повідомлень рекламного змісту) з підключенням створеної бази даних слів та словосполучень, характерних для декількох основних тематик кореспонденції (рекламні повідомлення, підписні розсилки-*newsletters*, персоналізовані повідомлення, несанкціоновані розсилки та ін.). Працюючи на основі баєсівського методу програма розподіляє вхідні повідомлення таким чином, що користувач, маючи можливість вносити зміни в налаштування, не відчуватиме незручностей, навіть не втрачаючи час на налаштування персоналізованих фільтрів.

Список літератури

1. Використання нейронної мережі кохонена для розпізнавання спаму – [Електронний ресурс]. – Режим доступу: http://pnzzi.kpi.ua/14/14_p106.pdf

2. Застосування методів навчання для розпізнавання спаму – [Електронний ресурс]. – Режим доступу: http://ena.lp.edu.ua:8080/bitstream/ntb/9546/1/15.pdf

3. Інтернет-реклама – [Електронний ресурс]. – Режим доступу: http://ru.wikipedia.org//wiki/Интернет-реклама

4. Терейковский И.А.Применение семантического анализа содержимого электронных писем в системах распознавания спама – [Електронний ресурс]. – Режим доступу: http://do.gendocs.ru/docs/index-206082.html

АНАЛІЗ ПРОБЛЕМИ АВТОМАТИЧНОГО ПОРОДЖЕННЯ АНГЛОМОВНИХ ДІЛОВИХ ЛИСТІВ

Волошина К. Ю.
Харківський національний університет
«Харківський політехнічний інститут»
м. Харків, ул. Пушкінська 79/1, тел. 80994377700,
e-mail: *kislaya.katia@yandex.ua*

В епоху розвитку інформаційних і комунікаційних технологій та їх впровадження в усі сфери людської діяльності проблема автоматичного породження текстів набула особливої актуальності. Це пов’язано у більшій мірі зі збільшенням кількості інформації, що подається у вигляді електронних документів. Актуальність сучасних досліджень в даній області пов'язана з розширенням і поглибленням ділових контактів як всередині країни, так і на міжнародному рівні. Таким чином, звернення до проблеми автоматичного породження англомовних текстів ділового характеру диктується потребою сучасного суспільства у створенні комп'ютерних систем, покликаних допомогти фахівцям, що працюють у сфері міжнародних і зовнішньоторговельних відносин. Подібні системи повинні значно скоротити час, необхідний на створення ділових листів, знизити ймовірність появи помилок, викликаних недостатньо високим рівнем володіння англійською мовою і технічними похибками, та знизити вартість створення англомовного ділового листа.

Результати дослідження сучасного стану справ в області розробки інтелектуальних комп'ютерних систем синтезу текстів на природній мові, свідчать про те, що їх побудова може бути здійснена в рамках двох принципово різних підходів:

1)системи, що працюють на основі шаблонних технологій, використовують готові репліки або комбінують готові фрагменти;

2)системи, що працюють на основі лінгвістично мотивованих технологій, призначені для створення текстів, що мають відносно вільний зміст, який не може бути заздалегідь заданим у вигляді фрагментів тексту[2].

Згідно з проведеному огляду, вдалося виділити найбільш відомі сучасні програми та сайти породження тексту: Scott Pakin's automatic complaint-letter generator, ANCHOR, Delirium 1.8, Generating the Web, SeoGenerator (SEO Anchor Generator), MonkeyWrite, Nice Letter (http://www.niceletter.com/).

Кожна з розглянутих програм, вимагає від користувача власноруч друкувати повністю тіло листа, зазначаючи його тематику та зміст. У програмах структура листа розбита на слоти, які необхідно заповнювати користувачу за винятком типових реквізитів, характерних для всіх типів листів а саме: «заголовок» листа, вказівка на посилання, дата, внутрішня адреса, вказівка на конкретну особу, вступне звертання, посилання на загальний зміст,

тема листа, заключна формула ввічливості, підпис, позначка про наявність додатків, позначка про надсилання копій на інші адреси, доповнення [1].

У результаті проведеного дослідження нами була запропонована наступна принципова схема алгоритму роботи автоматизованої програми породження ділових листів, яка представлена на рис.1.

Рисунок 1 - Алгоритм системи породження ділових електронних листів

Розглянемо кожний із етапів алгоритму створення англомовного ділового листа.

На I етапі користувач обирає одну із тематик листів.

Виходячи з того, що діловий англомовний лист має жорстку послідовну структуру, ми можемо використовувати шаблон структури листа, шаблонні вирази та характерні для кожної тематики листа речення із змінними даними, тому на II етапі створюється єдиний (початковий) шаблон, у якому у вигляді блоків описана структура майбутнього листа.

На III етапі користувач може включати або вилучати змінні варіанти полів з ключовими фразами, реченнями та словами, і тим самим може обрати один із багатьох варіантів закінчення та початку речення, а також, завдяки списку можливих варіантів підстановок більш індивідуалізувати свій лист. Також кожному листі наявні порожні поля які користувач має заповнювати особисто, наприклад: кількість товару, ім’я особи.

На IV етапі програма складає зміст усіх блоків структури та змінені та включені до листа поля у єдиний текст.

Подальше удосконалення такої системи автоматичного породження текстів можливо в декількох напрямках. Насамперед, це розширення меж застосування та використання даної технології для створення не тільки англомовних ділових документів, а й текстів, що належать до інших функціональних стилів і жанрів. Отримані дані можуть стати основою для створення багатомовних систем породження, що дозволяють синтезувати тексти на різних мовах.

Список літератури

1. *Коновченко О. В.* Міжнародне листування : навч. посіб. / – Х. : Нац. аерокосм. ун-т ім. М. Є. Жуковського «ХАІ», 2012. – 98 с.

2. *Соколова Е.Г.* Генерация текстов на естественном языке – состояние вопроса и прикладные системы / М.В. Болдасов, Е.Г. Соколова // НТИ. сер. 2. Информационные процессы и системы. –2005. – № 10. – С. 12–22.

МЕТОД ОПРЕДЕЛЕНИЯ КОРЕФЕРЕНТНЫХ СВЯЗЕЙ В МИНИМАЛЬНОЙ ЕДИНИЦЕ ДИСКУРСА

Терещенко В. И.
Национальный технический университет
«Харьковский политехнический институт»
г. Харьков, ул. Фрунзе, 21, тел. (057) 700-15-64
e-mail: omsroot@kpi.kharkov.ua

Одной из основных задач обработки естественного языка (Natural Language Processing) является установление связей между объектами в одном и том же тексте. Такие связи между элементами фраз в лингвистике называются кореферентными. На сегодняшний день существует достаточно много алгоритмов и моделей определения кореферентных связей в тексте. Однако большинство из данных разработок являются неприменимыми для флективных языков с хорошо развитой морфологией [1]. Хотя решение данной задачи является насущно-необходимым в приложениях Машинного перевода, Opinion Mining и автоматического реферирования, задача определения кореферентных связей в минимальной единице дискурса еще не была разрешена.

Кореферентность – отношение между местоимением (анафором) и его антецедентом, при котором и местоимение и его антецедент соотносятся с одним и тем же предметом объективной действительности [2]. Пример предложения с кореферентной связью:

*«Чистая прибыль Visa[**antecedent**] составила 1,27 млрд дол, говорится в сообщении компании. По итогам II квартала 2012 года она[**anaphor**] сообщила о чистой прибыли в 1,29 млрд дол.»*

Можно выделить следующие этапы разрешения данной задачи:

1) определение минимальной единицы дискурса;
2) выделение типов кореферентных отношений;
3) анализ текстов в которых встречаются кореферентные отношения;
4) разработка алгоритма определения кореферентных связей в минимальных единицах дискурса;
5) создание информационно-лингвистического обеспечения задачи определения кореферентных отношений в текстах заданного языка;
6) создание программной реализации разработанного алгоритма.

Дискурс (франц. discours – язык) – текст в некотором событийном аспекте; речь, которая рассматривается как целеустремленное социальное явление или компонент который берет участие во взаимодействии людей и механизмах их мышления (когнитивных процессах).

Выделяют несколько уровней письменного дискурса [3]:

1) синтаксический (фраза, предложение, высказывание);
2) лексический (токен, слово);
3) морфологический (морфема);
4) фонологический (фонема).

Следующим этапом исследования является выделение типов кореферентных связей. Текстам флективных языков наиболее присущи такие типы кореферентности [4]:

1) анафора;
2) катафора;
3) кореферентность именных групп;
4) сведение;
5) расширение.

В процессе исследования были определены особенности каждого из вышеперечисленных типов и для более глубокого исследования был выбран анафорический тип кореферентных связей, поскольку он встречается в более чем 80% проанализированных текстов. Этот тип кореферентности имеет структуру при которой антецедент всегда предшествует анафору в тексте. При этом анафор как правило является местоимением а антецедент существительным. Однако, исследования показывают, что не каждое местоимение является анафором и далеко не каждое существительное в тексте относится к анализируемому местоимению [5]. Взяв во внимание эти и другие установленные в ходе исследования особенности организации текстов на русском языке, было разработано множество грамматических правил для определения кореферентных связей в минимальных единицах дискурса. В результате чего к настоящему моменту было разработано 15 грамматических правил, включающих как синтаксические, так и морфологические закономерности анафорических связей в минимальной единице дискурса текстов русского языка. Был проанализирован массив текстов включающий более 200 предложений находящихся в свободном доступе на сайте GoogleNews.

Результатом исследования является программная реализация разработанная на основе алгоритма разрешения задачи определения кореферентных связей в минимальной единице дискурса учитывающего особенности русского языка. Программная реализация была выполнена средствами языка программирования Python 2.7. Работа приложения была протестирована на базе из 100 русскоязычных текстов общим объемом в 5,214Мб. Средняя точность автоматического определения анафоры приблизительно равна 69%.

Список литературы

1. *Clark J.H., Gonzalez-Brenes J.P*. “Coreference Resolution: Current Trends and Future Directions”, 2008. – 11-16p.

2. *Hobbs J. R.* “Pronoun resolution” – California, 1976. – 10-18p.

3. *Кашкин В.Б.* Сопоставительные исследования дискурса «Концептуальное пространство языка». Тамбов: ТГУ, 2005. С. 337-353.

4. *Скатов Д.* “Разрешение кореференции: обзорная экскурсия” – Н. Новгород, ДИКТУМ, 2012.

5. *Grosz B.J.* “Readings in natural language processing” – California, Morgan Kaufmann Publishers, Inc., 1986. - p. 339-352.

ВИКОРИСТАННЯ ГІПОНІМІЧНИХ ВІДНОШЕНЬ МІЖ ЕКОНОМІЧНИМИ ПОНЯТТЯМИ АНГЛІЙСЬКОЇ МОВИ ДЛЯ ПОБУДОВИ СЕМАНТИЧНИХ МЕРЕЖ

Булатнікова Т.С.
Національний Технічний Університет
«Харківський Політехнічний Інститут»
м. Харків, вул. Пушкінська, 79/2, тел. 0995692641,
e-mail: tbu2641@gmail.com

Метою даної роботи є розробка семантичних мереж гіпонімічних відношень економічних понять англійської мови, які можуть слугувати базі знань системи класифікацій існуючих компанії та корпорації за географічним положенням та галуззю виробництва.

Семантична мережа – це інформаційна модель предметної області, що має вигляд орієнтованого графа, вершини якого відповідають об'єктам предметної області, а ребра задають відношення між ними. Об'єктами можуть бути поняття, події, властивості та процеси.

Семантичні мережі відносяться до моделей класичного представлення знань у задачах штучного інтелекту, навчальних системах, системах машинного перекладу та семантичних павутинах. Підхід базується на трьох основних складових: суб'єкт, відношення, об'єкт. Саме ці три складові є базовими блоками семантичних мереж.

Мережеві моделі формально можна задати у вигляді $H =< I, C_1, C_2, ..., C_n, R >$, де I – множина інформаційних одиниць; C1, C2,..., Cn, – множина типів зв'язків між інформаційними одиницями. Відображення R задає між інформаційними одиницями, що входять до I, зв'язку із заданого набору типів зв'язків [1].

Більшість семантичних мереж базуються на ієрархічних відношеннях, які слугують для зв'язування мовних одиниць, що належать до різних рівнів. Фактично це піраміда, кожним рівнем якої керує більш високий рівень.

Вагома кількість таксономічних відношень також мають ієрархічну структуру. Таксономічні відношення або таксономія – це не що інше як класифікація. У ширшому сенсі таксономія стосується класифікації речей або понять, а також принципів, що лежать в основі такої класифікації, на відміну від меронімії, яка основана на класифікації частин цілого.

Використовуючи ієрархічну таксономію можливо послідовно ділити множину на підпорядковані підмножини, поступово конкретизуючи об'єкт класифікації. При цьому підставою для розподілу служить деяка вибрана ознака, а сукупність одержаних угруповань при цьому утворює ієрархічну деревоподібну структуру у вигляді графа, вузлами якого є угруповання [2].

Невід'ємною частиною поняття ієрархічної таксономії є гіпонімія. Гіпонімія як родо-видове відношення – сукупність семантично однорідних одиниць, які належать до одного класу. Гіпонімія характеризується привативною опози-

цією: видові назви завжди є семантично багатші від родових. Саме тому на відміну від синонімії, яка допускає взаємозаміну, гіпонімія характеризується односторонньою заміною гіпоніма на гіперонім, але не навпаки.

Гіпонімічні відношення – найбільш фундаментальні, парадигматичні і смислові, за їх допомоги структурується словниковий склад мови. На основі гіпонімії лексичні одиниці об'єднуються в тематичні й лексико-семантичні групи і поля. Саме тому, що панівними в лексико-семантичній системі є родо-видові відношення, превалюючим типом опозицій тут є інклюзивні, тобто відношення слабкого (немаркованого) і сильного (ознакового, маркованого) члена. Це надає лексико-семантичній системі домінантно-підпорядкованої впорядкованості (послідовне включення слів нижчого рівня абстракції до вищого), що не характерно для граматичних абстракцій.

Серед сфер використання гіпонімічних відношень не винятком є й економічна сфера. Гіпонімічні відношення у економіці – це система класифікації економічної діяльності, включаючи продукцію, компанії та галузі виробництва.

На сьогодні гіпонімія широко використовується для класифікації компаній за галуззю виробництва. Класифікація за галуззю виробництва впорядковує компанії у виробничі групи, що основане на схожості способів виробництва, продукції або поведінки на фінансовому ринку. Такі угрупування широко використовуються статистичними агенціями або у фінансовій сфері послуг для групування схожих інвестиційних компаній для створення індексів фінансового ринку за секторами [3].

Розроблені у дослідженні семантичні мережі гіпонімічних відношень мають вигляд ієрархії та класифікують компанії за географічним положенням головного офісу компанії та за галуззю виробництва, тобто мають таксономічну структуру та дозволяють класифікувати існуючі компанії та корпорації за географічним положенням та галуззю виробництва.

Побудова такої семантичної мережі, яка явним чином відображає родо-видові відношення географічного положення та галузі виробництва, дозволить розширити можливості систем автоматичної обробки англомовних текстів економічної тематики, пошукових систем, систем вилучення інформації, систем автоматичного реферування тощо, що в цілому дозволить отримувати більш конкретні данні з більш загальних і навпаки.

Таким чином побудовані у дослідженні семантичні мережі гіпонімічних відношень слугують не тільки для візуалізації впорядкованих даних, це також значний крок у сфері розвитку систем штучного інтелекту та інших систем, які базуються на знаннях.

Список литературы

1 Roussopoulos N.D. A semantic network model of data bases. – Department of Computer Science, University of Toronto, 1976. – p. 104.

2 Information intelligence: Content classification and enterprise taxonomy practice. Delphi Group. 2004. – p. 74

3 Day, A.C.L. The taxonomic approach to the study of economic policies. – The American Economic Review, 2010. – p. 78

THE INFLUENCE OF COHERENCE RELATIONS ON A SENTIMENT OF A DISCOURSE (BASED ON FRENCH NEWS ARTICLES)

Loda Sylvette
National Technical University "Kharkiv Polytechnic Institute",
Kharkiv, Pushkinskaya str., 79/2, tel. 707–63–60,
e-mail: simplement.sy@gmail.com

Coherence is currently a topic of intense debate in the international linguistic community, as well as automatic discourse analysis. The possibility of automatic semantic and sentiment analysis is very important in modern informational world. Now that we have an access to large amounts of information coming every day, it's becoming more and more complicated to analyze it.

Nowadays computational linguistics can operate various methods of sentiment and semantic discourse analysis, depending on the data they are applied to. But there is something that unites all of the topics and texts, here I am talking about connectives that are used to mark coherence relations between discourse segments. In my current research I want to address a question of a possible influence of connectives on semantic meaning and an overall sentiment of a text, narrowing the scope of research to financial news in French language as an example. Although it is a well-known fact that research into coherence strategy and automatic language processing is considered relevant to all spheres of human communication.

A text is coherent if it is designed around a common topic. In the reading process, the individual units of information enter meaningful relationships to one another. The text coheres and is not just a sequence of sentences to be interpreted in isolation. Those individual units that are united by specific relationships are called elementary discourse units (EDU). An EDU is a span of text, usually a clause, but in general ranging from minimally a Noun Phrase to maximally a sentence, that denotes a single event or type of event, serving as a complete, distinct unit of information that the subsequent discourse may connect to. This connection is performed by a certain coherence relation, which is a specific relationship, holding on the semantic or the pragmatic level of description, between adjacent units of text.

Coherence relations are usually explicitly marked with connectives, which belong to a closed-class non-inflectable lexical items. Their key property is their relational meaning: connectives set two discourse segments into correspondence with each other. From the semantic viewpoint, they therefore denote two-place relations. Syntactically, a connective can be a subordinating or coordinating conjunction, an adverbial, or (arguably) a preposition.

In my research I decided to define main classes of connectives with respect to their influence on a sentiment and semantic meaning of a discourse. The main idea was to make the classification of connectives as meaningful for practical use as possible. So basically we can identify the following groups of connectives according to their properties:

1 - connectives, changing sentiment of a nucleus sub-sentence and an overall sentiment of a phrase

E.g.: [L'abondance des sources alternatives de pétrole et les permis d'exploiter accordés en masse en Irak vont doper la production de pétrole]{S; pos}, mais [son prix restera élevé]{N, neg}. Overall sentiment: negative.

2 - connectives, influencing only satellite sub-sentence

E.g.: [Le dollar US a affiché son mois le plus bas depuis septembre]{N, neg} malgré que [Janet Yellen ait déclaré que la Banque Centrale va probablement maintenir sa politique monétaire]{S; pos}. Overall sentiment: negative.

3 - connectives that do not influence the sentiment of a span of discourse at all

E.g.: [Oddo réitère son opinion 'neutre' sur Aéroports de Paris]{pos} et [il relève l'objectif de cours à 88 euros]{pos}.

4 - connectives, that introduse a EDU giving different kinds of additional information (cause, consequence, result, time, place, example, comparison etc.)

E.g.: [Les ventes de camions légers ont grimpé de 9,7% en février, à 64 579]{pos}, alors qu'[elles avaient été de 58 867 en février 2013]{comparison}.

These groups are created out of a lexical base of French connectives which I observed in my previous course paper, containing 328 connectives. The new classification has a very important practical use in automatic discourse analysis and natural language processing, as it allows to enhance the effectiveness of sentiment and semantic analysis of texts.

Bibliography

1. Stede, Manfred (2011), Discourse Processing, Potsdam : Morgan & Claypool Publishers.

2. Charlotte Roze, Laurence Danlos (2009), Base lexicale des connecteurs discursifs du français, Paris: Université Paris Diderot.

3. Busquets, Joan (2013), Analyse du discours,Relations de cohérence, Bordeaux: Universit´e Bordeaux-3.

4. Bateman, John and Judy Delin (2005), « Rhetorical structure theory », In Encyclopedia of Language and Linguistics (2nd ed., pp. 589-596), Oxford : Elsevier.

5. Mann,William C. and Sandra A. Thompson (1986), Rhetorical Structure Theory : Description and Construction of Text Sructures (Technical Report No. ISI/RS- 86-174), Marina del Rey, CA : Information Sciences Institute.

6. Hobbs, Jerry, (1985). On the Coherence and Structure of Discourse (Research Report 85-37), Stanford, CA : CSLI.

7. Kehler, Andrew (2002), Coherence, Reference, and the Theory of Grammar, Stanford, CA : CSLI.

ПРИМЕНЕНИЕ МЕТОДА КОМПАРАТОРНОЙ ИДЕНТИФИКАЦИИ ДЛЯ ОБРАБОТКИ ЦИФРОВЫХ ИЗОБРАЖЕНИЙ ТЕПЛОТЕХНИЧЕСКИХ ПРОЦЕССОВ

Бабкова Н.В.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: nadjenna@gmail.com

Для последнего десятилетия характерны быстрое развитие технологий и методов компьютерной обработки цифровых изображений, а также появление скоростной цифровой фотоаппаратуры с высоким разрешением. Многие физические явления и величины, их характеризующие, являются по своей сути оптически наблюдаемыми. Традиционно при исследовании излучения при высокотемпературном процессе использовались пирометры. После внедрения в физический эксперимент компьютеров и цифровых фотокамер данный подход обрел вторую жизнь.

Методы цифровой обработки изображений находят применение в задачах анализа данных астрономических наблюдений, таких как реконструкция полей скорости солнечной плазмы, распознавание формы галактик и др. Такой подход позволяет получать не только качественную, но и количественную картину исследуемого физического процесса.

Как видим, оптические методы находят всё более широкое применение в различных физических экспериментах. Автоматизация данных исследований невозможна без использования цифровой фотосъемки в качестве средства измерения (или других методов позволяющих на выходе получить двумерное изображение), а методов цифровой обработки изображений — в качестве средства анализа результатов. Данный подход позволяет существенно снизить ресурсоемкость физического эксперимента, а также обеспечивает бесконтактное или псевдо бесконтактное измерение физических величин.

В связи со всем этим актуальной становится адаптация метода компараторной идентификации с использованием трехмерной модели цветового зрения человека для определения вида связи цветовой информации изображения и длиной волны (температурой) излучения с поверхности нагретого тела.

Если реальной системе сопоставить некоторый оператор, который описывает процесс ее функционирования, то на первый план выступает задача идентификации структуры этого оператора и его параметров. Довольно хорошо разработанная теория идентификации направлена на решение именно этого класса задач. В классическом понимании физическая постановка задачи идентификации неизвестного объекта выглядит следующим образом [3]: есть некоторый "черный ящик", т.е. объект или система, закономерности поведения которого мы хотим математически описать.

Теория психофизических процессов ставит своей задачей разработку математического описания зависимости ощущений от физических процессов, которые действуют на рецепторы человека. Формально компараторная идентификация может быть описана предикатом $E(x_1, x_2)$ вида [3]

$$E(x_1, x_2) = D(F[x_1], F[x_2]), \quad (1)$$

где x_1, x_2 – элементы множества входных сигналов B;

$y_1 = F[x_1],\ y_2 = F[x_2]$ – элементы множества выходящих сигналов B;

D – стандартный предикат равенства, заданный на декартовом квадрате множества B.

Для описания цветового восприятия предикат E принимает следующий вид:

$$E(b', b'') = D(f(b'(\lambda)), f(b''(\lambda))).$$

Здесь сигналы

$$f(b'(\lambda)) = u',\ f(b''(\lambda)) = u'', \quad (2)$$

где

$$u' = (u_i'),\quad u'' = (u_i'');\ i = 1,2,3; \text{ -} \quad (3)$$

трехмерные векторы с проекциями u_1', u_2', u_3' и u_1'', u_2'', u_3'', которые определяются по формулам

$$u_i' = \int_{\lambda_{1i}}^{\lambda_{2i}} b'(\lambda) K_i(\lambda) d\lambda;;\ u_{1i}'' = \int_{\lambda_{1i}}^{\lambda_{2i}} b''(\lambda) K_{1i}(\lambda) d\lambda;\ i = 1,2,3;.$$

Формулы (1) и (2) математически описывают вид функций $u' = f(b'(\lambda)),\ u'' = f(b''(\lambda))$.

D – предикат равенства, обусловленный таким образом:

$$D(u', u'') = \begin{cases} 1, & для \quad u' = u'', \\ 0, & для \quad u' \neq u''. \end{cases}$$

Данное описание представления предиката E легко интерпретируется в психофизических терминах. Сигналы u' и u'' можно понимать как цвета полей сравнения, субъективно воспринимаемые наблюдателем. Функция f – характеристическая функция эквивалентности - характеризует собой преобразование объективного светового излучения $b(\lambda)$ в субъективный цвет u, осуществляемое зрительной системой человека. Предикат D будем интерпретировать как операцию сравнения цветов, осуществляемую сознанием наблюдателя.

Рассмотрим трехмерную модель цветового зрения человека [3]. Зная значения функций спектральной чувствительности зрения $K_1(\lambda), K_2(\lambda), K_3(\lambda)$, которые представлены в табличном виде в [3], можно восстановить величину длины волны светового излучения.

Учитывая, что функции спектральной чувствительности зрения и излучение - числа, можно определить $K_1(\lambda), K_2(\lambda), K_3(\lambda)$:

$$K_i = \frac{u_i}{b(\lambda_2 - \lambda_1)},\ i = 1,2,3;$$

По цифровому изображению высокотемпературного процесса определим координаты цвета *R, G, B* в компьютерной системе цветовых координат. С помощью формул перехода определим координаты цвета u_1, u_2, u_3 в системе МКО. Для определения величины светового излучения b используем метод компараторной идентификации.

Для иллюстрации разработанного способа измерения приведем результаты обработки фотографии горящей парафиновой свечки (рис. 1). Для рис. 1 на изображении пламени зафиксировано значение максимальной относительной температуры 871 K и коэффициента теплопроводности $\lambda = 50,9$ Вт/(м·К).

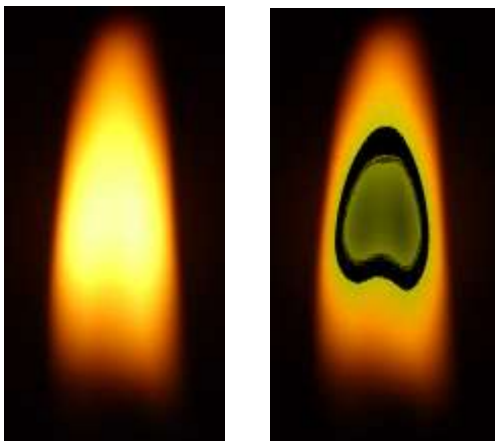


Рис. 1 − Горение парафиновой свечи

Выводы. Привлечение современных средств регистрации изображений, методов цифровой обработки изображений, распознавания образов позволяет осуществлять высокоточное количественное исследование оптически наблюдаемых или специально визуализированных физических процессов. Обработка больших массивов данных, полученных в ходе экспериментов, может быть в значительной степени автоматизирована. Это позволяет исследователю экономить время, затрачиваемое на проведение анализа и интерпретацию данных, а также на идентификацию изучаемых процессов на основе моделей, построенных для их описания.

Список литературы

1. *Вавилов В.П.* Тепловые методы неразрушающего контроля: Справочник –М. : “Машиностроение”. − 1991. − 240с.

2. *Смирнов А.Д.* Математические модели теории передачи изображений –М. : “Связь”. − 1979. − 96 с.

3. *Бондаренко М. Ф.* Мозгоподобные структуры: Справочное пособие. / М. Ф. Бондаренко, Ю. П. Шабанов-Кушнаренко. Том первый. Под редакцией акад. НАН Украины И.В. Сергиенко. – К. : Наукова думка, 2011. – 460 с.

4. *Бондарев В. Н.* Искусственный интеллект / В. Н. Бондарев. – Севастополь : СевНТУ, 2002. – 615 с.

РАЗРАБОТКА АЛГОРИТМА ВЫДЕЛЕНИЯ ЭЛЕМЕНТОВ ЭМОЦИОНАЛЬНОСТИ

Медведская А.В.
Национальный технический университет
«Харьковский политехнический институт»
61002, Харьков, ул. Фрунзе, 21, тел. (057) 707-64-60
e-mail: brovka90@inbox.ru

С развитием информационных технологий и всемирной паутины Интернет формируется Интернет язык, который, по форме существования является письменным языком, фактически близок к разговорному, поскольку обладает ее основными признаками: непосредственностью и неподготовленностью общения, преобладанием диалога над монологом, эмоциональностью, экспрессивностью и логической непоследовательностью высказываний.

Таким образом, если большая часть коммуникации переходит в область письма, неизбежно встает вопрос о недостаточности письменной речи. Во время речи мы используем различные средства передачи эмоций: мимику, интонацию, жестикуляцию. На письме мы лишены этих возможностей. Но в языке интернета для передачи эмоций существуют некоторые инструменты, которые помогают понять эмоции пользователя. Следовательно, происходит обогащение языка [1].

Эти средства можно классифицировать следующим образом:

1. Фонетические:

•многократное повторение звуков (пролонгация);
•написание по слогам (скандирование или произношение);
•искривление стандартов орфографии;
•словесное ударение с помощью большой гласной буквы;
•фразовый ударение;
•интонационная окраска.

2. Графические. Основные графические средства выражения эмоций – это графические улыбки (смайлики). Широкой популярности в форумах, чатах они достигли из-за удобства вставки в текст, понятности, возможности выражения эмоций.

3. Для выражения эмоций существуют различные лексические средства, которые уже по своему содержанию эмоциональные, выражающие то или иное чувство [2, 3].

Целью данной работы является решение задачи определение эмоциональной лексики в электронных сообщениях. Эта задача зачастую является одной из подзадач анализа тональности текста.

Для решения задачи определения эмоциональных элементов используются следующие методы: методы, основанные на знаниях (правилах, словарях); Machine Learning и скрытая Марковская модель.

Для построения алгоритма извлечения эмоциональных элементов из электронных сообщений был выбран метод Machine Learning с учителем.

Общий алгоритм Machine Learning выглядит следующим образом:

1) Необходимо собрать коллекцию документов для обучения классификатора.

2) Каждый документ учебной коллекции нужно представить в виде вектора признаков. Сообщение с учебной коллекции нужно разметить тегами.

1. Многократное повторение звуков (проголгация).

<prolong>аааааааа</prolong>

2. Написание по слогам (скандирование или произношение).

<scans> Поз-дра-вл яю</scans>

3. Искривление стандартов орфографии.

<fault>Аффтар</fault>

4. Словесное ударение с помощью большой гласной буквы.

<wstress> КрасавЕц </wstress>

5. Фразовое ударение.

<phstress> ДА </phstress>

6. Интонационная окраска.

<tone>!!!!</tone>

7. Смайлик.

<smile>:)</smile>

3) Для каждого документа надо указать «правильный ответ», то есть, содержит эмоциональные элементы или не содержит, за этими ответами и будет обучаться классификатор.

4) С помощью классификатора система начинает сравнивать два класса, к которым мы отнесли сообщения на предыдущем этапе, и обнаруживает характерные признаки эмоциональности в тексте. Так система учится на тренировочной коллекции.

5) Используя тестовую выборку, тестируем обученный классификатор. Документ подается на вход классификатора, и полученные ответы сравниваются с экспертными оценками [4].

В ходе работы был разработан алгоритм определения эмоциональных элементов в электронных сообщениях с использованием Machine Learning, с помощью которого система будет обучаться распознавать и выделять эмоциональные элементы в тексте.

Список литературы

1. *Валгина Н. С.* Активные процессы в современном русском языке: учебное пособие для студентов вузов /Н. С. Валгина. – Москва: Логос, 2001. – 210 с.

2. *Голуб И. Б.* Стилистика русского языка. /И. Б. Голубь – М.: Айрис-пресс, 2003. – 155 с.

3. *Максимов В. И.* Стилистика и литературное редактирование: учебник /В. И. Максимов. – М.: Гарда-рики, 2007. – 274 с.

4. *Пак А.* Обучаем компьютер чувствам (sentiment analysis по-русски) [Электронный ресурс] – Режим доступа: http://habrahabr.ru/post/149605/

АЛГОРИТМ АВТОМАТИЗИРОВАННОГО РЕФЕРИРОВАНИЯ НОВОСТНЫХ АНГЛОЯЗЫЧНЫХ ТЕКСТОВ

Дашкевич Е.С.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 0978314534,
e-mail: esdashkevich@gmail.com

Целью данного исследования является разработка математической модели автоматизированного реферирования текстов на английском языке. В процессе выполнения работы были поставлена задача разработать математическую модель построения автоматического реферата и на основе разработанной модели построить алгоритм программной реализации автоматического реферирования.

Процесс разработки методов автоматического формирования краткого представления или реферата (англ. summary) текстовых документов длится с конца 1950-х годов. Автоматическое реферирование – это создание коротких выкладок материалов, аннотаций или дайджестов, т.е. получение важнейших сведений из одного или нескольких документов, и генерация на их основе лаконичных и информационно-насыщенных отчетов [1]. Существуют два направления автоматического реферирования - квазиреферирование и краткое изложение содержания. Современные методы реферирования дают возможность автоматизировать данный процесс таким образом, что вмешательство человека и конечная корректировка продукта будут минимизированы. На сегодняшний день пользователю доступны системы, осуществляющие автоматическое реферирование разными методами и на разных языках, и работающие с текстами разных тематик.

Информация, которую человек получает ежедневно – это новости. Наибольшим спросом, по мнению новостного агентства БиБиСи, пользуются сжатые новостные сообщения аналитического характера. Основным методом, который используется для реферирования новостных сообщений, является статистический. Это обусловлено тем, что новостные сообщения представляют собой четко структурированные тексты, не требующие специальной семантической обработки.

Учитывая, что язык – это конечный, дискретный, детерминированный объект, для своего моделирования он требует методов и средств дискретной математики. Дискретная математика позволяет моделировать большинство языковых явлений и анализировать языковые процессы.

Для расчёта веса термина используется формула TF. Как правило, наибольший вес в документе имеют общеупотребительные слова и термины, которые не дают представления о содержании текста. Термины со слишком малым весом также могут быть ключевыми за редким исключением, поэтому должны быть исключены из текста [2]. С целью получения ключевых слов по данному принципу, будет использовано равенство веса ключевых терминов

$TF_{min} < TF_k < TF_{max}$, где TF_{min} – нижняя граница веса терминов в документе, TF_{max} – верхняя граница веса терминов в документе, TF_k – диапазон веса ключевых слов. Таким образом, задача автоматического реферирования состоит в том, чтобы создать реферат, максимально приближенный по качеству к получаемому в результате человеческой когнитивной деятельности.

На основе этого утверждения был создан алгоритм реферирования англоязычных новостных сообщений, который опирается на метод веса ключевых слов. Суть алгоритма заключается в том, чтобы выделить из текста лексемы со средним весом и считать их ключевыми. Для удобства работы с текстом, он будет приведен к массиву строк. На стадии предварительной обработки будет определен объем текста. Так мы получим обработанный текст, благоприятный для дальнейшего анализа.

В полученном «скелете» текста будет проведен анализ количества вхождений конкретного термина в текст. Для этого все лексемы, начиная с первого слова в заголовке, будут сравниваться со следующими лексемами в массиве предложений. Найденная лексема сразу удаляется из исходного предложения и заносится в структуру, где ей присваивается значение инкременты 1. Так, количество вхождений термина в текст инкрементируется начиная со значения 1++. Посредством удаления уже обработанных лексем осуществляется оптимизация алгоритма. Вследствие этого этапа получим необходимые параметры для определения веса терминов. После вычета веса всех терминов, анализатор определит нижнюю и верхнюю границу веса ключевых терминов. Все лишние лексемы будут удалены из структуры. Последним этапом автоматизированного создания реферата является запись заголовка текста и всех предложений, содержащих ключевые слова, в текстовый файл. Таким образом, получим реферат новостного сообщения.

Данный алгоритм также может работать с корпусом текстов. Основой создания реферата корпуса текстов является нахождение ключевых слов, общих для всех текстов данной коллекции. Благодаря выделению цельных ключевых предложений в реферат, минимизируется возможность нарушения синтаксических связей и дробления смысла и идеи текста. Цель заключается в том, чтобы сохранить оригинальную идею текста и использованные в нем семантические единицы.

Таким образом, в результате работы представлена математическая модель автоматизированного реферирования новостных текстов. Был предложен алгоритм для программной реализации системы автоматизированного реферирования, основанный на методе веса ключевых слов. Следующим шагом в решении задачи автоматизированного реферирования является создание соответствующего программного обеспечения.

Список литературы

1. *Nenkova A., McKeown K.* Automatic Summarization./ A. Nenkova, K. McKeown. – NY. : Springer US, – 2011. – pp. 216.

2. *Sparck Jones K.* A statistical interpretation of term specificity and its application in retrieval./ K. Sparck Jones. – L.: Journal of Documentation, –1972. – pp. 12.

ПРОБЛЕМЫ РАСПАРАЛЛЕЛИВАНИЯ ПРОЦЕССОВ ПРИ ОПРЕДЕЛЕНИИ АВТОРСТВА ТЕКСТОВ

Груздо И.В., Россоха С.В., Шостак И.В.
Национальный аэрокосмический университет
им. Н.Е. Жуковского «ХАИ»
г. Харьков, ул. Чкалова 17,
e-mail: tigralwovna@rambler.ru, sergey.rossokha@gmail.com,
Iv_shostak@rambler.ru

Современный этап развития сферы образования в Украине характеризуется, с одной стороны, стремительным ростом объемов письменных работ, а с другой – расширением их номенклатуры. Проверка этих работ нуждается в тщательном анализе со стороны преподавателей с целью установления оригинальности контента, а также в одновременном использовании при оценивании работ четырех разнотипных систем. Указанные обстоятельства порождают проблему, суть которой состоит в недостаточной эффективности существующих, «ручных» способов обработки преподавателями текстов письменных учебных работ (ПУР), установлении фактов наличия в них текстовых заимствований, и формировании на основании этих фактов объективных оценок работ в четырех разнотипных системах оценивания.

В докладе рассмотрена методика поиска заимствований в ПУР. Отличием описанной методики анализа текстовых документов является учет особенностей, присущих письменным учебным работам, что, в конечном итоге, приводит к повышению эффективности организации учебного процесса в заведениях различного уровня аккредитации, а так же выбор соответствующего метода анализа текстов в зависимости от типа ПУР.

На основе данной методики была разработана специальной ИТ анализа и оценивания письменных учебных работ с учетом наличия в них текстовых заимствований, а так же был рассмотрен процесс анализа ПУР в техническом университете, а именно, Национальном аэрокосмическом университете им. Н.Е.Жуковского «ХАИ». При этом автоматизация анализа ПУР была осуществлена с использованием ПС «Plagiarizm», который представляет собой исследовательский прототип, функционирующий на основе специально разработанной технологии.

Использование ПС «Plagiarizm» в учебном процессе дает следующий эффект: накопление опыта преподавателей и нормоконтролеров, а также для поддержки принятия решений ими по оцениванию письменных учебных работ. Кроме того, следует отметить, что применение ПС «Plagiarizm» на практике предоставляет возможность оценивать эффективность результатов, достигнутых в ходе исследования.

Для того, чтобы использовать преимущества ПС «Plagiarizm» необходимо выполнить преобразование последовательно выполняемых процессов программы в несколько параллельно взаимодействующих процессов. При этом необхо-

димо, что бы машина автоматически выполняла однозначное определение отдельных файлов для каждого обнаруженного позаимствованного объекта, так же связывала полученную информацию по каждой проанализированной работе, с соответствующим файлом источником, и производила анализ полученных данных. Для этого была выполненная формулировка задачи распараллеливания процессов при определении авторства текстов и рассмотрены существующие типы распараллеливания.

В работе были проанализированы существующие архитектуры многопоточных приложений:

– многопроцессные приложения с автономными процессами;

– многопроцессные приложения, взаимодействующие через трубы, сокеты и очереди System V IPC;

– многопроцессные приложения, взаимодействующие через разделяемую память;

– собственно многопоточные приложения;

– событийно-ориентированные приложения;

– гибридные архитектуры.

В ходе анализа был сделан вывод о том, что по своей направленности ПС «Plagiarizm» относится к событийно-ориентированным приложениям, а так же имеет сложно организованную структуру, когда в составе программы имеются и параллельные участки и циклические. При преобразовании ПС «Plagiarizm» в параллельную программу необходимо гарантировать, что все обработчики событий будут завершаться достаточно быстро.

Было отмечено, что важное место в преобразовании ПС «Plagiarizm» в параллельную программу занимают теоретические вопросы, включающие выбор модели распараллеливания, определение эквивалентности программ, определение степени параллелизма и синхронизация параллельных процессов, анализ потенциального параллелизма алгоритмов и программ, определение максимального параллелизма и др.

В докладе было приведено описание параллельного алгоритма поиска заимствований и определения авторства текстов, а также разработанная группа алгоритмов, позволяющих за приемлемое время получать распределения параллельных ветвей анализа текстов ПУР по процессорным ядрам. Первая группа алгоритмов относится к классу вероятностных алгоритмов локальной оптимизации, вторая основана на алгоритмах обхода графов и их разбиении.

ЗАДАЧА КЛАСИФІКАЦІЇ ТЕКСТІВ АНГЛІЙСЬКОЮ МОВОЮ ЗА ГЕНДЕРНИМИ ОЗНАКАМИ

Борзенкова А.В.
*Національний технічний університет
"Харківський політехнічний інститут",
м. Харків, вул. Пушкінська, 79/2, тел. 707–63–60,
e-mail: borzenkova-alina@yandex.ru*

На сьогоднішній день можна говорити про існування гендерних досліджень, що вивчають обидві статі, а точніше – процес соціального конструювання розходжень між статями. Гендер вважається соціокультурним конструктом, пов'язаним із приписуванням індивіду певних якостей і норм поведінки на основі його біологічної статі [1].

Соціальні науки вступили в епоху науки даних, використовуючи безпрецедентні джерела писемності [2-4]. Через засоби масової інформації, такі як Facebook і Twitter [5], регулярно користуються більше 1/7-й населення світу, простежуються відмінності між жіночим та чоловічим мовленням. Щоб розібратися у масивних даних, необхідні багатопрофільні співробітництва між такими областями, як комп'ютерна лінгвістика та соціальні науки. У даній роботі демонструється інструмент, який описує схожості та відмінності між групами людей з точки зору їх використання мови.

У самому загальному плані дослідження гендера у мовознавстві стосується двох груп проблем.

1. Мова і відображення в ньому статі. Мета такого підходу полягає в описі і поясненні того, як маніфестується у мові наявність людей різної статі (досліджуються в першу чергу номінативна система, лексикон, синтаксис, категорія роду та ін.), які оцінки приписуються чоловікам і жінкам і в яких семантичних областях вони найбільш помітно та чітко виражені.

2. Мовну, і в цілому комунікативну, поведінку чоловіків і жінок, де виділяються типові стратегії і тактики, гендерно специфічний вибір одиниць лексикону, способи досягнення успіху у комунікації, переваги у виборі лексики, синтаксичних конструкцій та ін. – тобто специфіка чоловічого і жіночого мовлення [6].

Актуальність цього дослідження обумовлюється важливістю визначення, вивчення і опису способів формування гендерних стереотипів, які несвідомо і/або усвідомлено закладаються у світосприйнятті людиною за допомогою сприйняття їм текстів повідомлення, впливу з боку засобів масової інформації та безпосередньої комунікації в соціумі. Важливість дослідження обумовлена збільшеною потребою лінгвістів, культурологів, психологів, педагогів в освоєнні механізмів, що надаються інтернет-простором для самопрезентації особистості, а також в адекватному і детальному визначенні тієї ролі, яку сьогодні відіграє віртуальний світ у житті людини, надаючи йому варіант альтернативної реальності. Самопрезентація у соціальній мережі є для сучасної людини одним з найбільш важливих атрибутів мовного оформлення та підтвердження самобу-

тності власного Я.

Метою роботи є вирішення задачі класифікації текстових повідомлень за гендерними ознаками.

Задача класифікації – формалізована задача, в якій є множини об'єктів (ситуацій), розділених деяким чином на класи. Задана кінцева множина об'єктів, для яких відомо, до яких класів вони відносяться. Ця множина називається вибіркою. Класова приналежність інших об'єктів невідома. Необхідно побудувати алгоритм, здатний класифікувати довільний об'єкт з вихідної множини.

У процесі роботи над роботою зроблено огляд основних напрямків гендерної лінгвістики, методів класифікації текстових документів та існуючих систем, що реалізовують класифікацію текстових документів. Автором були проаналізовані особливості гендерних ознак у соціальних мережах та розроблено алгоритм класифікації текстів з використанням методу опорних векторів.

Метод опорних векторів виявляє закономірності в даних і створює структури, які можуть бути використані для класифікації текстових повідомлень. Першим кроком є підготовка вибірки даних, які будуть використані для навчання нашої моделі для розпізнавання повідомлень методом опорних векторів. Навчальна вибірка представляє собою набір вхідних даних і відповідних їм вихідних даних, які використовуються для аналізу та вилучення патерну [7].

Виходячи з результатів лінгвістичних досліджень, можна зробити висновок про те, що у висловлюваннях чоловіків і жінок дійсно існують помітні відмінності. При цьому вони можуть бути представлені у формі, допустимою для комп'ютерної обробки, а значить, їх можна використовувати і в методах машинного навчання для класифікації текстів.

Інші характеристики особистості автора, за даними лінгвістів, також впливають на мову людини. Тому, використовуючи цю інформацію при класифікації, можна сподіватися на отримання більш точних даних про автора анонімного тексту.

Список літератури

1. Coats J. Women, men and language. A sociolinguistic account of sex differences in language. – New York, 1986. – 389 p.

2. Lazer D, Pentland A, Adamic L, Aral S, Barabasi AL, et al. (2009) Computational social science. – Science 323: 721-723.

3. Weinberger S (2011) Web of war: Can computational social science help to prevent or win wars? the pentagon is betting millions of dollars on the hope that it will. Nature 471: 566–568. doi: 10.1038/471566a

4. Miller G (2011) Social scientists wade into the tweet stream. Science 333: 1814–1815. doi: 10.1126/science.333.6051.1814

5. Facebook (2012) Facebook company info: Fact sheet website. Available:http://newsroom·fb·com. Accessed 2012 Dec.

6. Preisler B. Linguistic Sex Roles in Conversation. – Paris: Mauton, 1986. – 287 p.

7. Ageev M.S., Dobrov B.V. Support Vector Machine Parameter Optimization for Text Categorization Problems. // Вестник Национального Технического Университета «ХПИ» – Харьков, Украина, 2004. – №1 – 3-14 с.

ВИКОРИСТАННЯ СУЧАСНИХ КОМП'ЮТЕРНИХ ТЕХНОЛОГІЙ ПРИ НАВЧАННІ ГРАМАТИКИ АНГЛІЙСЬКОЇ МОВИ

Кудоярова О. В.
Національний аерокосмічний університет імені М.Є.Жуковського «Харківський авіаційний університет»
м. Харків, вул. Чкалова 17, тел. 788-48-71
e-mail: olga_kudoyarova@mail.ru

У процесі інтеграції України до європейської спільноти та потреба суспільства розпочати реформування системи вищої освіти зумовлює розробку та впровадження нових методик, які б сприяли професійному становленню майбутнього фахівця та його майстерності. Змінені умови навчання потребують від викладача використання більш ефективних прийомів та методів контролю і оцінки знань, умінь та навичок студентів. Для підвищення рівня володіння іноземною мовою студентами, та формування граматичних навичок необхідно покращити ефективність контролю за її вивченням.

В сучасній педагогічній практиці все частіше використовується метод тестового контролю навчально-пізнавальної діяльності студентів, який базується на єдиній та об'єктивній методиці виявлення, оцінки та обліку навчальних досягнень студентів. Сьогодні комп'ютерні тести є одним з електронних засобів навчання та перевірки знань.

На заняттях з граматики англійської мови зі студентами денної форми навчання використання методик тестового контролю дозволяє виконувати контролюючу функцію навчального процесу, яка з одного боку підтримує оперативний зворотній зв'язок для цілеспрямованого керування процесом навчання студентів, а с другого боку визначає ефективність методик та прийомів навчання, які використовує викладач; навчальну функцію, яка сприяє керуванню навчанням, формуванню вмінь та навичок, їх коректуванню та вдосконаленню; діагностико-керуючу функцію, яка дозволяє виявити рівень оволодіння елементами знання (вмінь та навичок) та розглядає результати навчання в залежності від шляхів та способів їх досягнення; стимулююче-мотиваційну функцію, яка сприяє підвищенню мотивації пізнавальної діяльності, інтересу до предмету та стимулює студентів покращити свої результати; розвиваючу функцію, яка дозволяє розвивати пам'ять, увагу, логічне мислення, формують самостійне творче мислення, уміння робити висновки, узагальнювати, використовувати знання в нових ситуаціях. Найважливішими критеріями якості тесту є його валідність, об'єктивність, надійність і точність.

Методики тестового контролю використовуються як для денної так і для дистанційної форми навчання.

Посередником між студентом і викладачем при дистанційній формі навчання стає система дистанційного навчання, яка є цілим комплексом модулів, що відповідають за окремі етапи навчання. До складу систем дистанційного навчання входять: подача теоретичного матеріалу (лекція, мультимедія, опис

практичної/лабораторної роботи, глосарій,бібліотека), перевірка засвоєння подачі матеріалу (звіт по практичній/лабораторній роботі, питання/завдання до, спілкування студентів з викладачем і між собою (питання до лекції off-line, ICQ, Skype, E-mail, телефон, форум тощо**).** Сьогодні існує велика кількість систем дистанційного навчання. В практиці дистанційного навчання багато вищих навчальних закладів використовують систему дистанційного навчання Moodle.

Moodle – це система управління вмістом сайту (Content Management System), спеціально розроблена для створення якісних онлайн-курсів викладачами. Moodle орієнтована на колаборативні технології навчання – дозволяє організувати навчання в процесі спільного вирішення навчальних завдань, здійснювати взаємообмін знаннями. Moodle перекладена на десятки мов, в тому числі і російську, і використовується у 197 країнах світу. Модульна структура системи забезпечує простоту використання системи для студентів і викладачів. Перевагами системи дистанційного навчання Moodle є: широкий набір можливостей для повноцінної реалізації процесу дистанційного навчання: різні опції формування і представлення навчального матеріалу, перевірки знань і контролю успішності, спілкування і організація студентського співтовариства; активне залучання студентів в процес формування знання і їх взаємодію між собою; потужні системи глосарію і форуму; можливість реалізувати проекти різних рівнів складності; багатомовний інтерфейс; велика географія поширення по всьому світу; програмне забезпечення з відкритими початковими кодами під ліцензією GPL, яке надає можливість безкоштовного використання системи.

На кафедрі прикладної лінгвістики розроблені навчальні матеріали з граматики англійської мови, які складаються з теоретичного та практичного матеріалу для системи дистанційного навчання Moodle. На базі системи було створено банк питань, граматичні довідки, тестові завдання. Граматичні довідки та відеоматеріали, які допоможуть студентам виконувати дані тести, представлені перед кожним тестом в елементі курсу «семінар».

Розроблені тестові завдання з граматики англійської мови та використання системи дистанційного навчання Moodle дозволять розв’язати низку нових дидактичних завдань, а їх застосування забезпечить підвищення якості освіти.

Отже, розвиток інформаційного суспільства та впровадження комп’ютерних технологій суттєво впливають на модернізацію всіх сфер діяльності. Зокрема, якісна система підготовки фахівців вимагає пошуку засобів, нових методик і ефективних технологій освітньої діяльності.

МЕТОДИ СТВОРЕННЯ НОВОГО УНІКАЛЬНОГО ТЕКСТОВОГО КОНТЕНТУ САЙТУ

Тупікова Н.С.
Національний технічний університет
«Харківський політехнічний інститут»
61002, Харків, вул. Фрунзе, 21, тел. (057) 707-64-60
e-mail: elpys@rambler.ru

На зорі появи Інтернет-технологій не потрібно було докладати багато зусиль для відвідуваності сайту. Їх було не так багато. Але через час конкуренція зросла – в мережі з'явилася велика кількість ресурсів.

Успішність і розкрутка сайту в пошукових системах, залежать від ряду факторів. Дуже важливим фактором є контент. Контент – збірний термін для будь-якої інформації, яка міститься в інформаційному ресурсі. Тексти та зображення, що розміщуються на сайті, мають бути унікальними. Унікальним контентом може називатися будь-який контент, який ще жодного разу не був опублікований в мережі. Унікальність контенту – є одним з ключових моментів у просуванні сайту, оскільки пошукові машини при індексації визначають, чи був цей текст використаний раніше на інших веб-ресурсах [2].

Підвищення унікальності є досить складним завданням. Зазвичай контент роблять унікальним за допомогою рерайтингу. Рерайтинг – це процес створення нового унікального тексту з вже існуючого тексту. Рерайт – це продукт, який отримуємо на виході цього процесу. Зовні рерайтинг нагадує написання шкільного переказу – передача суті розповіді своїми словами [5]. Систему такого переписування вже багато разів намагалися автоматизувати. Програм і онлайн – сервісів, існує безліч, серед них – програми підбору синонімів [4].

Синонімайзер – програма або сервіс для заміни слів у тексті на синоніми, що знаходяться в базі даних з метою видозміни тексту і надання йому унікальності. Синонімайзери можна розділити на автоматичні і ручні. Ручні пропонують користувачеві самому вибирати зі списку синоніми, які підходять. Автоматичні проробляють всю роботу без участі користувача. Синонімайзери поділяються на серверні або он-лайнові, тобто ті, що працюють виключно в Інтернеті, та десктопні, ті, що встановлюються на комп'ютер користувача.

Серед он-лайнових синонімайзерів відомі synonym.savenkoff.name, synonymizer.ru, synonyma.ru. До десктопних синонімайзерів відносяться: Smartrewriter, Synmonster, USyn, WordSyn, DimSyn [6].

Існують методи, за допомогою яких, змінивши текст, можна отримати новий унікальний текст.

1. Синоніми. Цей метод займає перше місце. Його суть – заміна слів синонімами.

2. Заміна дієслів іменниками. Його суть зводиться до заміни дієслів відповідними іменниками.

3. Заміна прямої мови непрямою. Кожен текст може містити цитати. Оскі-

льки якісний рерайт не передбачає прямої мови, її краще замінити непрямою.

4. Маніпуляції з реченнями. Цей метод передбачає зміну структури речень: поділ складного речення на два або більше простих, або навпаки – об'єднання двох або більше простих речень в одне складне.

5. Використання пасивного стану. Цей метод передбачає перестановку місцями присудків [3].

Таким чином, результатом рерайтинга повинен виявитися повністю унікальний текст, що містить в собі знання, ідеї, думки і факти, які вже існують.

У даній роботі для моделі створення унікального текстового контенту був узятий принцип роботи синонімайзера, а також метод заміни прямої мови непрямою.

Дія синонімайзера заснована на механізації роботи з текстами. Якість роботи синонімайзера безпосередньо залежить від словника синонімів, який використовує програма. Відбувається проста заміна X на Y за допомогою словника виду X | Y.

Для заміни прямої мови непрямою було проаналізовано два види речень.

Якщо непряма мова передає повідомлення, розповідь, то речення, яке передає чужу мову, приєднується сполучником «що».

Сашко сказав: «Я нікуди не піду». – Сашко сказав, що він нікуди не піде.

Існують також випадки, коли пряма мова не стосується автора або не відноситься до нього, тобто з’являється деякий суб’єкт.

Батько сказав: «Проект сподобався». – Батько сказав, що проект сподобався.

Якщо непряма мова передає питання, то речення, приєднується сполучниками «що», «хто», «який», «де», «коли», «куди», «звідки».

Дарина запитала: «Куди я поклала олівець?». – Дарина запитала, куди вона поклала олівець.

Розглянемо випадок, коли з’являється інший суб’єкт.

Хлопчик запитав: «А де море?» – Хлопчик запитав, де море. [1].

Список літератури

1. Заміна прямої мови непрямою: [Електронний ресурс]. – Режим доступу: http://school.xvatit.com/index.php?title=Тема_2._Прямая_и_косвенная_речь_как_способы_передачи_чужой_речи

2. Контент сайту: [Електронний ресурс]. – Режим доступу: http://igroup.com.ua/seo-articles/kontent/

3. Методи створення унікального тексту : [Електронний ресурс]. – Режим доступу: http://shard-copywriting.ru/copywriting-basics/rewrite-rewriting-examples

4. Програми для рерайтера: [Електронний ресурс]. – Режим доступу: http://mastersloga.ru/news/programmy_dlja_rerajtera/2013-04-14-47

5. Рерайтинг: [Електронний ресурс]. – Режим доступу: http://igroup.com.ua/seo-articles/rerajtynh/

6. Синонімайзер: [Електронний ресурс]. – Режим доступу: http://animatika.ru/info/gloss/synonymizer.html

АНАЛІЗ ДЕРИВАЦІЙНИХ МОДЕЛЕЙ В УКРАЇНСЬКІЙ МОВІ

Стребкова О.О.
Національний технічний університет
«Харківський політехнічний інститут»
г. Харків, вул. Фрунзе, 21, тел. (057) 707-63-60,
e-mail: strebkova@ro.ru

Вічним двигуном у мові є словотвір, який діє без перепочинку щохвилини, щосекунди. Двигун словотвору діяв, діє і діятиме повсякчас, поки з'являтимуться нові предмети, народжуватимуться нові слова і думки.

Слово – дуже важлива мовна одиниця. Воно широке у побудові своєї форми, глибоко наповнене змістом та гнучке у своїх різноманітних функціях. Слово не ізольоване від інших одиниць, як і від іншого мовного контексту, і не замкнене в собі. Всі слова взаємопов'язані і за формою, і за змістом та співвідносяться між собою [2].

Словотвір, або дериватологія (від лат. derivacio – відхилення, утворення), – це розділ мовознавства, який вивчає закони утворення похідних слів від інших спільнокореневих слів. Словотвір став однією зі самостійних галузей українського мовознавства у 60-х рр. XX століття. Засновником дериватологічної школи вважають І. Ковалика, який у своїх працях пояснив теоретичний апарат словотвору, визначив його принципи та тенденції [3].

Словотвір як розділ мовознавства пов'язан з іншими розділами української мови, такими як фонетика, фонологія, лексикологія, морфологія й синтаксис. Розрізняють морфологічні й неморфологічні способи словотворення. До морфологічних відносять всі способи, за яких похідне слово утворилося за допомоги афіксальних морфем, всі інші – до неморфологічних [2].

Морфологічні способи словотворення поділяються на афіксальні, безафіксальні (осново- і словоскладання, абревіація). Залежно від того, які афікси використовуються для творення похідних слів, серед афіксального словотворення виокремлюють такі його різновиди: суфіксальний, префіксальний, суфіксально-префіксальний, постфіксальний, безафіксний.

До неморфологічних способів словотворення належать: морфолого-синтаксичний, лексико-синтаксичний і лексико-семантичний [2-6].

Проблема спілкування з комп'ютерами на природній мові породжує широке коло завдань, вирішення яких передбачає координацію зусиль математиків, лінгвістів, програмістів. Одне з цих завдань полягає у розробці моделей мови, що дозволяють здійснити розпізнавання машиною значень адресованих їй повідомлень. Однією з найважливіших проблем при створенні систем обробки природномовної інформації є адекватна інтерпретація значень повідомлень, вирішення якої неможливе без розуміння значення слова [1].

У наш час вивчення процесу словотворення становить інтерес не тільки з погляду на те, як організована мова для вираження значення, але й з погляду розширення можливостей й оптимізації роботи систем автоматичної обробки

інформації. Включення в них моделей словотворчого аналізу дозволяє здійснювати обробку потенційних слів і неологізмів і разом з тим значно скоротити обсяг машинного словника системи. Крім цього словотворчий аналіз істотно полегшує виявлення та корекцію помилок у вхідному тексті. Все це з очевидністю свідчить про необхідність вивчення й моделювання словотворчої системи мови.

Метою даної роботи – дослідження і розробка системи автоматичного префіксального словотворення дієслів української мови.

Для моделювання семантики української мови, що характеризується величезною різноманітністю похідних форм, велике значення має вивчення і формалізація опису словотвірної семантики. Як відомо, зміст похідного слова в загальному випадку можна описати як просту суму значень складових його морфів. Для того щоб краще зрозуміти механізм формування значення деривата, необхідно дослідити та описати математичні міжморфемні семантичні зв'язки, що мають місце в процесі деривації між префіксальними і кореневими морфемами, кореневими та суфіксальними морфемами, а також між основами і закінченнями. Моделювання словотворення можливо описати у вигляді системи рівнянь алгебри скінченних предикатів[1].

Надамо опис моделі міжморфемних семантичних відносин на множині префіксальних ланцюжків дієслів української мови. Розглянемо дві множини морфем які задані предикатами $P_1\,(t_1)\,\epsilon\,S_1$ та $P_2\,(t_2)\epsilon\,S_2$, які характеризуються узгодженням певними семантичними ролями цих морфем. У результаті семантичного узгодження двох морфем які стоять поруч отримуємо множину зв'язків між семантичними ролями, іншими словами, – множину пар семантичних ролей. Таким чином, між множинами семантичних ролей морфем які стоять поруч існує бінарне відношення, яке є підмножиною декартового множення цих множин. Це бінарне відношення можна представити за допомогою двомісного предиката $P(t_1, t_2)$, при цьому:

$$P(t_1, t_2) \rightarrow P_1(t_1) \cdot P_2(t_2) \qquad (1)$$

Деякі семантичні ролі для морфем які стоять поруч насправді не вступають в узгодження, у зв'язку з цим у формулу (1) вводиться додатковий множник $\lambda\,(t_1, t_2)$, який прагне виключити нереалізовані зв'язки. Таким чином, бінарне відношення на множинах для морфем які стоять поруч може бути задане формулою (2):

$$P_1\,(t_1) * P_2(t_2) = \lambda\,(t_1, t_2) \cdot P_1\,(t_1) \cdot P_2(t_2), \qquad (2)$$

Логічний добуток предикатів $(P_1\,(t_1) \cdot P_2(t_2)$ формули описує всі можливі зв’язки між морфемами, а предикат $\lambda\,(t_1, t_2)$ виключає частину нереалізованих зв’язків. Для множини префіксів це означає наступне. Серед префіксальних ланцюжків переважну більшість яких становлять двопрефіксні ланцюжки, що мають вид Π_1 та Π_2, де Π_1 – префікс, що складається на першому місці в ланцюжку, Π_2 – на другому. Як приклад:

Розглянемо багатозначні дієслівні префікси ПРИ- і ПО- , що стоять на першому місці в префіксальної ланцюжку, а також префікси ВІД- і ПО-, що стоять на другому місці.

Префікс ПРИ- має наступні семантичні ролі:

X_1 – «за допомогою названої дії досягнути будь-якого місця» (прибігти, приїхати);

X_2 – «названою дією зблизити, з'єднати що-небудь» (прив'язати, приклеїти);

Для префікса ПО- характерні наступні семантичні ролі:

X_8 – «початок дії» (побігти, поїхати);

X_9 – «здійснити дію для прийому їжі» (поїсти);

Далі, семантичні ролі морфеми ВІД- , що стоїть на другому місці в префіксальному ланцюжку, мають наступний вигляд:

Y_1 – «відділення одного предмета від іншого» (відірвати, відколоти);

Y_2 – «здійснити у відповідь або зворотню дію» (віддати);

Префіксальна морфема ПО- має наступний набір семантичних ролей:

Y_6 – «за допомогою названої дії досягнути будь-якого місця» (повзти);

Y_7 – «тривалість дії» (поїздити);

Аналіз словника показав, що префіксальна пара ПРИ- ВІД- реалізує наступні композиції семантичних ролей:

X_7Y_1 – «неповнота відділення одного предмета від іншого» (привідкрити);

Для префіксальної пари ПРИ- ПО- має наступні композиції семантичних ролей:

X_1Y_6 – «за допомогою названої дії досягнути будь-якого місця» (приповзти);

Префіксальна пара ПО- ВІД- має наступні композиції семантичних ролей:

$X_{11}Y_1$ – «поширення дії на багато предметів і відокремлення одного предмета від іншого» (повідривати);

І, нарешті, префіксальна пара ПО- ПО- має наступні значення:

X_8Y_7 – «тривалість дії протягом деякого часу» (попоїздити);

В ході роботи була розроблена модель префіксального словотворення дієслів у вигляді системи рівнянь алгебри скінченних предикатів, що пов'язують приставки з наборами, які однозначно характеризують їх ознаки.

Список літератури

1. *Шаронова Н. В.* Компараторная идентификация лингвистических объектов: дис. … док. тех. наук : 05.25.05 / Н. В. Шаронова. – Харьков, 1994. – 271 с.

2. *Кочерган М. П.* Вступ до мовознавства. Підручник. – Вид. 2-ге. – К., 2008 – 368 с.

3. *Дорошенко С. І.* Загальне мовознавство. Підручник. – Вид. Київ, 2006 – 289 с.

4. Словотвір: [Електронний ресурс]. – Режим доступу: http://pidruchniki.ws/1292052240545/dokumentoznavstvo/slovotvir

5. Предмет словотворення: [Електронний ресурс]. – Режим доступу: http://distribut.net/article/a-240.html

6. Словотвір як окремий розділ мовознавства: [Електронний ресурс]. – Режим доступу: http://lessons.com.ua/slovotvir-yak-okremij-rozdil-movoznavstva-osnovni-ponyattya-slovotvoru-metodika-vivchennya-slovotvoru-v-shkoli/

СТВОРЕННЯ НАВЧАЛЬНОГО ДОВІДНИКА ВІДМІНЮВАННЯ ДІЄСЛОВА У ФРАНЦУЗЬКІЙ МОВІ

Переваруха С.Г.
Національний технічний університет
"Харківський політехнічний інститут",
м. Харків, вул. Пушкінська, 79/2, тел. 707–63–60,
e-mail: sofia.perevarukha@mail.ru

Активне застосування комп'ютерних технологій відкриває доступ до нових джерел інформації, активізує навчально-пізнавальну діяльність, скорочує час вивчення мови, дає нові можливості для розвитку професійних навичок та їх вдосконалення, значно підвищує ефективність самостійної роботи. Тому надзвичайно важною стає розробка електронних засобів навчання зокрема у сфері іноземних мов.

При вивченні французької мови найбільших труднощів, на наш погляд, викликає засвоєння системи дієвідмінювання, оскільки вона є досить розгалуженою та має велику кількість винятків. Тому основною метою даної роботи стало створення довідника, який направлений на відпрацювання навичок відмінювання дієслів французької мови.

У процесі роботи над теоретичною частиною було зроблено огляд існуючих програм [4, 5]з вивчення французької мови, які можна умовно розділити за наступними напрямками:

- для засвоєння нових слів;
- для тренування французької вимови;
- для вивчення граматики;
- для загального підвищення рівня знання французької мови та інші.

Але серед знайдених програм, присвячених вивченню системи дієвідмінювання французької мови, більшість була лише готовими довідниками [4], а не навчальними програмами [5], у яких можна виконати вправи на закріплення навичок дієвідмінювання. Таким чином було зроблено висновок, що створення електронного навчального довідника з дієвідмінювання французької мови є актуальною задачею.

Для досягнення поставленої мети потрібно було вирішити наступні завдання:

1) побудувати алгоритм роботи електронного довідника;
2) створити базу даних закінчень та основ неправильних дієслів;
3) розробити інтерфейс та програмне забезпечення.

Алгоритм програми було засновано на граматиці французької мови. За основу було взято традиційний розподіл дієслів на три групи [1, 2] та розподіл за кількістю основ кожної групи [3]. У повному вираженні алгоритм є достатньо об’ємним, тому пропонується виділити 4 етапи основні етапи. На 1-ому етапі користувач обирає для дієвідмінювання дієслово будь-якої групи в інфінітиві. На 2-ому етапі програма виконує пошук цього слова у базі даних дієслів, ви-

значає групу, до якої воно відноситься, та відокремлює основу, а закінчення дієслова в інфінітиві відкидається. На 3-му етапі користувачеві пропонується дописати відповідне закінчення у кожній особі однини та множини до відокремленої основи. На 4-ому етапі програма робить перевірку відповідей, введених користувачем. У результаті програма міститиме різні вправи за складністю та типом, наприклад, для тренування дієслів 3-ої групи доцільно перевіряти не знання закінчень, а того, як змінюється основа. Тим не менш, загальна ідея залишається тією ж: проводимо опис закінчень дієслів та групуємо дієслова за типами основ у базі даних дієслів.

База даних навчального довідника складається з двох таблиць, куди внесено основи неправильних дієслів 2-ої та 3-ої групи та закінчення. Ці дві таблиці зв'язані між собою за особою та кількістю. Окремо створено таблицю неправильних дієслів 3 групи, які не мають правил дієвідмінювання.

Інтерфейс програми буде представлений наступним чином: програма буде мати 3 вкладки. Перша – «Довідка», де користувач вводе інфінітив дієслова, та програма його відмінює та виводе форми слова. У наступній вкладці користувач може перевірити свої знання з дієвідмінювання 1 та 2 групи дієслів, під час виконання вправ, де програма надає особу та основу дієслова, а користувач має ввести закінчення та суфікси. В третій вкладці користувачеві надається лише інфінітив дієслова та особи, а користувач має ввести повністю дієслово. Окрім того для створення граматичних асоціацій на сторінці виводиться список дієслів, які відмінюються за тими же принципами, що і задане дієслово. Тобто при невірному виконані вправи користувач зможе закріпити свої знання на інших дієсловах.

Таким чином, у ході роботи було описано особливості та ознаки кожної групи дієслів, які лягли в основу словозмінної парадигми дієвідмінювання французької мови у теперішньому часі. На основі цих особливостей та зроблених висновків після проведеного огляду існуючих програм було розроблено алгоритм для створення навчального довідника дієвідмінювання французької мови, будується база даних закінчень та розробляється програмне забезпечення.

Список літератури

1. *Опацький С. Є.* Français, niveau débutant: Підруч. для вищ. навч. закл. / Опацький С. Є. – Київ: «Перун», 2005. – 312 с.

2. *Попова И. Н., Казакова Ж. А., Ковальчук Г. М.* Французский язык. Учебник для I курса институтов и факультетов иностранных языков: учеб. пособие /Попова И. Н., Казакова Ж. А., Ковальчук Г. М. – Москва: «Нестор Академик», 1997. – 576 с.

3. *Christine Grall* Grammaire progressive du français/ Christine Grall – Paris: CLE International, 2003. – 192 p.

4. Довідник відмінювання французьких дієслів [Електронний ресурс] – Режим доступу: www.leconjugueur.com

5. Довідник відмінювання французьких дієслів з вправами [Електронний ресурс] – Режим доступу: www.bonjourdefrance.com

ИСПОЛЬЗОВАНИЕ ЛИНГВИСТИЧЕСКИХ КОРПУСОВ ПРИ ОБУЧЕНИИ ИНОСТРАННЫМ ЯЗЫКАМ: ПЛЮСЫ И МИНУСЫ

Лесная М. И.
Харьковский национальный университет им. В. Н. Каразина
г.Харьков, пл. Свободы, 4, тел. 7075136,
e-mail: marianna1983@yandex.ru

В рамках современной образовательной парадигмы основной целью обучения иностранным языкам является формирование коммуникативной компетенции, что предполагает активное использование аутентичных дидактических материалов. Кроме неадаптированной литературы, аудио- и видео- информации общего и специализированного характера, которые традиционно применяются на уроках иностранного языка, в педагогической практике все активнее используются также новые методы, основанные на данных лингвистических корпусов. Этот подход соответствует концепции data-driven learning, или DDL [5, с. 295], в рамках которой учащиеся получают возможность анализировать язык «изнутри», используя реальные и актуальные лингвистические данные, становятся активными исследователями, создателями и соавторами языковых правил и закономерностей, развивая свою самостоятельность и сознательность в процессе обучения.

Как отмечает П. В. Сысоев, формирование лексико-грамматических навыков на основе использования лингвистического корпуса возможно исключительно в рамках проблемного подхода, который позволяет активизировать речемыслительную деятельность учащихся, вследствие чего полученные знания хорошо и надолго усваиваются [1, с. 105-106].

Стоит отметить, что большинство изучающих иностранные языки уже неосознанно пользуются наработками корпусной лингвистики, поскольку современные грамматики, учебные курсы и словари, например, издательств Cambridge, Collins, MacMillian, Oxford, Pearson Longman и др. основаны на объективных и актуальных данных, полученных из лингвистических корпусов. Эта информация применяется для анализа наиболее употребительных слов, фраз и грамматических конструкций, отслеживания изменений в семантике лексем, определения, какое значение должно быть заглавным в словарной статье, отображения грамматических и синтагматических характеристик языковых единиц и т.п.

Что касается сознательного использования корпусных данных, оно может быть прямым (hands-on) или опосредованным (hands-off) [3, с. 550]. Прямой подход предполагает работу учащихся с корпусным программным обеспечением, что может повлечь ряд сложностей. Во-первых, аудитории должны быть обеспечены достаточным количеством компьютеров, подключенных к сети Интернет. Во-вторых, необходимо учитывать вероятность нестабильного соединения и сбоев в работе корпусного менеджера. В-третьих, определенное время на занятии будет потрачено на объяснение принципов работы с корпусами, интер-

фейсы которых не всегда отличаются дружественностью, а также на решение сопутствующих технических проблем. В-четвертых, нужно быть готовым к отсутствию энтузиазма у той части ученической аудитории, которые склонны к более традиционным формам работы.

Кроме того, преподаватель сталкивается с вопросом методологического характера: насколько адекватно привлечение корпусных данных для решения той или иной дидактической задачи? Исходя из нашего опыта, использование указанного типа информации нецелесообразно на начальных уровнях овладения языком (A1-A2 согласно общеевропейской шкале оценивания уровня языковой компетенции). Проблема заключается в том, что корпусная выдача в основном неоднородна и может не соответствовать допустимому уровню сложности. Также корпус не всегда дает релевантные дидактической цели данные, и преподавателю зачастую необходимо специально отбирать наиболее показательные примеры.

На наш взгляд, прямое использование лингвистических корпусов при изучении иностранного языка оптимально при анализе и исправлении ошибок, поскольку в этом случае учащиеся имеют четко сформулированную цель, достаточную мотивацию найти ответ на конкретный вопрос, а также средства реализации поставленной задачи, а именно доступ к большому количеству примеров достоверного и аутентичного словоупотребления. Также имеет смысл формировать у студентов навык использования лингвистического корпуса в качестве справочного ресурса в ходе выполнения творческих письменных заданий. Роль учителя в этом случае состоит в том, чтобы воздержаться от прямых ответов на вопросы «А можно ли так сказать?» или «Какое слово лучше употребить?» и помочь студенту найти нужную информацию в корпусе.

При опосредованном подходе учащиеся не работают с самим лингвистическим корпусом, а имеют дело с материалами, заранее подготовленными преподавателем на основе корпусных данных. Исследователи предлагают различные варианты заданий, создаваемых с привлечением корпуса, большинство из которых сводятся к следующим:

1) определение значения языковой единицы на основе анализа ближайшего контекста;

2) изучение полисемии слова, поиск новых значений лексем;

3) составление словарной статьи на основе корпусной выборки;

4) изучение сочетаемости лексем;

5) выявление различий в употреблении близких по значению слов;

6) анализ особенностей использования грамматических структур, формулирование правил их употребления [1, 2, 4, 6, 7 и др].

Как показывает практика, задание 1) не работает, если языковая единица, которую предлагается исследовать, является абсолютно незнакомой. Анализ разнородных контекстов, которые к тому же являются только фрагментами предложений, зачастую не позволяет «схватить» смысл нового слова или словосочетания и в основном вызывает негативную реакцию учащихся. Дело обстоит иначе, если языковая единица уже встречалось ранее, есть некое пред-

ставление о ее семантике, которое необходимо уточнить. В этом случае привлечение лингвистического корпуса является абсолютно оправданным.

Что касается заданий 2) и 3), они, на наш взгляд, не способствуют формированию коммуникативной компетенции учащихся и больше подходят для семинарских занятий по лексикологии или лексикографии.

Задания 4) и 5) особо эффективны на продвинутых уровнях владения языком, когда студенты имеют достаточный словарный запас, но делают ошибки при употреблении слов и словосочетаний в контексте.

Формулирование правил использования грамматических конструкций на основании анализа примеров их употребления (задание 6), по нашему мнению, является оптимальным способом обучения грамматике, особенно если имеется возможность представить данные контексты на более ранней стадии занятия в виде целостного текста или аудио-материала.

Подводя итог выше сказанному, стоит отметить, что применение корпусных методов в лингводидактике имеет неоспоримые преимущества (в частности, при подготовке словарей, грамматик и учебных курсов), но также характеризуется рядом особенностей: они особо продуктивны на продвинутых уровнях языковой компетенции, у мотивированных студентов, которые расположены к проблемно-исследовательской деятельности.

Список литературы

1. *Сысоев П. В.* Лингвистический корпус в методике обучения иностранным языкам / П. В. Сысоев // Язык и культура. – 2010. – №1. – С. 99–111.

2. *Bernardini S.* Corpora in the classroom: An overview and some reflections on future developments / Silvia Bernandini // How to use corpora in language teaching. – Amsterdam, 2004. – Pp. 15-36.

3. *Boulton A.* Data-driven learning: Taking the computer out of the equation / Alex Boulton // Language Learning. – Volume 60. – Issue 3. – Pp. 534-572.

4. *Corpora and language learners* / edited by Aston G., Bernardini S., Stewart, D. – Amsterdam/Philadelphia: John Benjamins Publishing, 2004. – 311 p.

5. *Johns T. F.* From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning / T. F. Johnes // Perspectives on Pedagogical Grammar. New York: Cambridge University Press, 1994. – Pp. 293-313.

6. *O'Keeffe A.* From Corpus to Classroom / Anne O'Keeffe, Michael McCarthy, Ronald Carter. – Cambridge: Cambridge University Press, 2007. – 315 p.

7. *Reppen R.* Using Corpora in the Language Classroom / Randi Reppen. New York: Cambridge University Press, 2010. – 118 p.

ДЛЯ НОТАТОК

Наукове видання

Матеріали

III Всеукраїнської науково-практичної конференції

"ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ ТА ПРИКЛАДНА ЛІНГВІСТИКА"

(Українською, російською та англійською мовами)

Відповідальний за випуск *Н.В. Шаронова*

Технічна редакція та комп'ютерна верстка: *С.В. Петрасова*

Формат 60х90/16. Ум. друк. арк. 4,9. Наклад 50 прим.

Надруковано у ТОВ «Планета-Принт»
61002, м. Харків, вул. Фрунзе, 16. Свідоцтво № 24800170000040432 від 21.03.2011 р